语/言/与/认/知/译/丛

MINDS, BRAINS, COMPUTERS:
AN HISTORICAL INTRODUCTION TO THE FOUNDATIONS OF COGNITIVE SCIENCE

心智、大脑与计算机：
认知科学创立史导论

◎〔美〕R. M. 哈尼什　著
王　淼　李鹏鑫　译



MINDS, BRAINS, COMPUTERS:
AN HISTORICAL INTRODUCTION TO THE FOUNDATIONS OF COGNITIVE SCIENCE

本书从历史和跨学科两个视角，对认知科学基础问题进行了导论性质的述评。它描绘了从19世纪到当代认知科学领域某些核心概念的发展轨迹，考察了哲学、心理学、神经科学和计算机科学等学科对认知科学发展的重要贡献和影响。该书还从两个形成鲜明对照的进路考察了心智理论问题：数字计算机和神经网络模型。该书内容可靠、全面，是认知科学导论课程的理想读本，也是心智哲学课程极好的补充读物。

这是一部引人入胜的著作，它为认知科学的基本观念提供了颇富魅力的、论述翔实的历史导论。本书不仅对于研习认知科学的学生（将从大量训练中获益）非常重要，而且对于想要获取认知科学内自己从事领域之外其他领域知识的学者，以及想对这门学科的发展历史有深入了解的人，都会从中获益。

——美国马里兰大学哲学系Georges Rey教授

在认知科学课程中，有两个问题令人苦恼：（1）来自某个学科的学生缺乏对其他学科背景的足够了解，而这些学科对于他所学专业来说是至关重要的；（2）即便在自己所学的学科之中，学生往往不能拥有理解本学科当代关注问题的来源和重要性所必需的历史观点。哈尼什的这本内容丰富、见闻广博的著作专门为解决这两个问题而作，并且获得了极大的成功。

——美国罗格斯大学哲学系Stephen Stich教授


ISBN 978-7-308-07273-1
9 787308 072731
定价：65.00元

YR 语/言/与/认/知/译/丛　　黄华新　盛晓明　主编

MINDS, BRAINS, COMPUTERS:
AN HISTORICAL INTRODUCTION TO THE FOUNDATIONS OF COGNITIVE SCIENCE

心智、大脑与计算机：

认知科学创立史导论

◎〔美〕R. M. 哈尼什　著
王　淼　李鹏鑫　译



图书在版编目（CIP）数据

心智、大脑与计算机：认知科学创立史导论 /（美）R. M. 哈尼什著；王淼，李鹏鑫译. —杭州：浙江大学出版社，2010. 8

Minds, Brains, Computers: An Historical Introduction to the Foundations of Cognitive Science

ISBN 978-7-308-07273-1

Ⅰ. ①心… Ⅱ. ①哈… ②王… ③李… Ⅲ. ①认知科学—研究 Ⅳ. ①B842. 1-09

中国版本图书馆 CIP 数据核字（2010）第 160502 号

浙江省版权局著作权合同登记图字：11—2008—52 号

This edition is published by arrangement with **Blackwell Publishing Ltd**, Oxford. Translated by **Zhejiang University Press** from the original English language version. Responsibility of the accuracy of the translation rests solely with **Zhejiang University Press** and is not the responsibility of **Blackwell Publishing Ltd.**

心智、大脑与计算机：认知科学创立史导论

Minds, Brains, Computers: An Historical Introduction to the Foundations of Cognitive Science

[美] R. M. 哈尼什 著　王　淼　李鹏鑫 译

策　　划　曾建林
责任编辑　田　华
封面设计　刘依群
出版发行　浙江大学出版社
（杭州市天目山路 148 号　邮政编码 310007）
（网址：http://www.zjupress.com）
排　　版　杭州中大图文设计有限公司
印　　刷　德清县第二印刷厂
开　　本　710mm×1000mm　1/16
印　　张　27.75
字　　数　500 千
版 印 次　2010 年 8 月第 1 版　2010 年 8 月第 1 次印刷
书　　号　ISBN 978-7-308-07273-1
定　　价　65.00 元

“语言与认知译丛”总序

人类的心智(mind)和行为也许是宇宙间最顶端、最复杂也是最奇异的现象了,但人类只有通过自身的心智和行为才能认识和理解自己。无怪乎美国著名的认知神经科学家达玛西奥(A. Damasio)在研究意识时发出这样的感叹:“还有什么比知道如何知道更困难的事情呢?正因为我们有意识,才使我们能够,甚至不可避免地要对意识提出疑问,还有什么比认识到这一点更让人惊异和迷乱的呢?”“知道如何知道”——这正是认知科学的根本任务,而且也是促使其从哲学认识论中萌芽并最终在当代的哲学—科学研究中枝繁叶茂的根本动力。

认知研究已成为当前世界大国国家科技战略特别关注的领域之一。一个日益普遍的看法是:对心智的科学认识将在人类认识自身、科学技术、医学发展、经济增长、社会安全、人类幸福和生活品质的提高等人类和国家利益方面产生革命性的影响!世界众多一流大学或相应机构都在这个领域进行着你追我赶的研究,力图率先取得原创性的成果;加强和促进认知科学的发展同样符合我国的国家科技战略目标。《国家中长期(2006—2020年)科学和技术发展规划纲要》将“脑科学和认知科学”列为8个基础前沿研究领域之一,而且加快了对认知科学的资助和研究机构的规划部署。自“985工程”一期和二期实施以来,相继有一些高等院校和科研院所建立了以认知研究为重点的研究机构。浙江大学语言与认知研究中心(CSLC)就是“985工程”二期面向认知研究的人文社会科学与自然科学兼容的哲学社会科学创新基地之一。

认知科学有“一个长的过去,但只有一个相对短的历史”。也许正因为其历史短暂,其发展态势就显得尤为迅捷。自20世纪50年代“认知革命”发

生以来，认知科学一直处于高速发展的阶段。图中列出的一些重要的学术事件清楚地展示了这一点。面对这种情势，CSLC自项目启动伊始就怀有强烈的紧迫感。然而另一方面，当前认知科学的研究局面斑驳陆离，这是历史上任何一个学科在其发展中都不曾有过的。至今认知科学还没有一个公认的统一的学科边界，还处在统一范式形成的前夜：研究的基本观念、维度、问题域和方法都复杂多样。为了在这个驳杂的局面中明确定位，形成特色，我们认为必须对当前认知研究的格局和趋势有一个较为全面的认识，从而根据自己的优势，在权衡慎思后提出自己的问题并开展深度研究，为推动认知科学在我国的发展尽自己的职责。基于这个考量，CSLC决定选译一些认知研究著作，作为系列丛书连续出版。对选译的著作，CSLC的设想非常简明：(1)根据CSLC文理兼容、偏向哲学社会科学的研究特色，选译著作应有很强的思想性；(2)这些著作的思想观念不求经典，但却是开拓新研究方向，融合新研究方法的创始之作。此动议萌生之时，CSLC就开始着手选题和组织翻译，历时两年余，"语言与认知译丛"首批作品开始陆续奉献于读者面前。译事辛苦，尽管各书译者都勤勤恳恳，几易其稿，但不足乃至错讹之处可能仍难避免，诚恳期望学界同仁和广大读者朋友批评指正。在此成书之际，CSLC尤其感谢浙江大学出版社的真情投入和热情支持。

加速发展的认知科学

CSLC"语言与认知译丛"主编

黄华新　盛晓明

2008年9月

中文版序：认知科学与心智计算理论概述

在《心智、大脑与计算机》中文版出版之际，我愿意对书中详细探讨的有关心智计算理论的观点进行概述。

“心智计算理论”（computational theory of mind, CTM）在帕斯卡（Blaise Pascal，1623—1662）、霍布斯（Thomas Hobbes，1588—1679）以及莱布尼茨（Gottfried Wilhelm von Leibniz，1646—1716）等人的著作中就有所论及，但它的真正繁荣是在巴贝奇（Charles Babbage，1791—1871）十进制机器、楚泽（Konrad Zuse，1910—1995）二进制机器（约 1940 年）以及图灵（Alan Mathison Turing，1912—1954）抽象计算机（约 1936/1937 年）等开创性工作基础上而出现于 20 世纪后半叶。事实上，CTM 的繁荣也伴随着认知科学在 20 世纪六七十年代的诞生和发展。

学界通常对认知科学存在两种理解，而这两种理解都与 CTM 密切相关。将认知科学看作是研究认知的交叉学科，我称之为**广义认知科学**——正如这个名称所表达的——本书图 I. 1 即为它的图解。但是很多学者对这种理解并不满意，认为认知科学并非仅是这些有关认知研究的学科总称——他们期望一种更为具体的能够涵括这种研究取向的定义。因此也就有了**狭义认知科学**，将认知看作是一种系列计算，也就是 CTM。

但倘若认知最终被证明并非是计算的，那么对认知科学的这种理解恐怕会不幸地失去其研究主题，而这似乎是难以接受的。最好的办法是允许并且接受认知科学是“研究认知的交叉学科”这样的广义理解，同时也接受对认知科学的狭义理解，即目前最好的认知理论也就是将认知看作是一种系列计算。那么就有了如下三个问题：

Q1：认知是计算，这种观点如何产生？

Q2：这种观点究竟确切地说明了什么？

Q3：这种观点存在什么难题，它具有怎样的前景？

Ⅰ. CTM 的历史:从亚里士多德到图灵

CTM 的出现受到很多智识传统和学科的重要影响。这里,我将简要回顾从亚里士多德到图灵对于 CTM 产生所作的一些贡献。这部分内容主要有两个主题:一个与可内省的常识心理学相关;另一个与非可内省的非常识生物学相关。我将前者称为“软件”主题,而将后者称为“硬件”主题。

亚里士多德　在我看来,这两个主题均来自亚里士多德(Aristotle,公元前 384—前 322):

软件主题:亚里士多德对记忆和回忆的讨论,都是基于记忆内容的联想链接;

硬件主题:亚里士多德将脑看作是血液的散热器——思维由心脏产生(因此人们说“用心学习”)。

A. 截止到图灵:软件主题

(i)联想主义　17—19 世纪,英国经验主义者霍布斯、洛克(John Locke,1632—1704)、休谟(David Hume,1711—1776)、密尔(James Mill,1773—1836)等对亚里士多德有关记忆联想结构的讨论,作了最为突出的探讨。本书图 1.1 标识了经验联想主义关于心灵观念的一般观点。联想主义主要持有如下假设:

(1)心灵通过联想原则由简单观念构成复杂观念(同样参考图 1.1);

(2)观念基于联想原则随时间相互联结(见图 1.2)。

联想主义一直是心理学的主导理论,直到 20 世纪由于下列因素的出现导致了联想主义走向衰落:

(1)人们对内省主义方法的强烈不满;

(2)脑神经科学研究的快速发展(见下文);

(3)“科学”心理学观念的出现(见下文)。

(ii)行为主义　在当时人们看来,更为“科学”的心理学观念应该是行为主义。纽约《时代周刊》(1942)曾宣称行为主义是“探索人类心智的新时代”。它起始于华生(John Broadus Watson,1878—1958)《行为主义者心目中的心理学》(1913),该文被评为《心理学评论》出版史上最为重要的一篇文献。华生在文中所讨论的一些主要问题成为后来行为主义的核心观点,直到 1960 年以后行为主义的衰落。他的主要观点如下:

(1)心理学是一门客观的自然科学;

(2)心理学的任务是预测和控制行为；

(3)人类与动物心理是连续的，并无本质不同；

(4)内省和意识不属于心理学的研究内容。

华生的观点在后来45年里得到著名心理学家赫尔(Clark Leonard Hull，1884—1952)、奥斯古德(Charles Egerton Osgood，1916—1991)、托尔曼(Edward Chase Tolman，1886—1959)等人的深入研究。最为极端的是斯金纳(Burrhus Frederic Skinner，1904—1990)，他在《语言行为》(1957)中尝试将他早年动物操作条件("斯金纳箱")的工作扩展到人类语言领域。

(iii)认知主义 1959年，年轻的语言学家乔姆斯基(Chomsky，1928—)出版了一部与斯金纳观点针锋相对的著作，被历史学家称之为"华生行为主义宣言之后最有影响力的心理学著作"。最重要的是，乔姆斯基进一步提出在解释行为时需要考虑心灵具有其本身的主动操作原则的观点。

在米勒(George Armitage Miller，1920—)、加兰特尔(Eugene Galanter)与普利布拉姆(Karl H. Pribram)的《行为的规划和结构》(1960)中贯彻了这种观点，被普遍认为开启了认知主义对行为主义的回应。该著作提出了一种TOTE(Test-Operation-Test-Exit，测试一操作一测试一输出)单元，用以替代行为主义的"反射弧"，如本书图2.2所示。

这些单元具有如下主要特征：

(1)单元包含着有机体关于世界的信息表征；

(2)涉及一种控制单元序列操作的系统；

(3)它们能够以无限的方式彼此嵌套。

下面是分别与计算机关键概念的心理学类比：

(1)表征知识的数据结构；

(2)转换控制程序；

(3)有助于提高编程效率及灵活性的子程序。

B. 截止到图灵：硬件主题

亚里士多德的"硬件"猜想在历史上并没有获得支持，但有关"硬件"的理解在亚里士多德之后并没有立即走向正确，而是先变得更加离奇。

预览　神经科学的历史，在很大程度上可以看作是尝试寻找脑的心理功能与其结构间映射的历史，因此长期伴随着定位论和整体论两种取向之间的争论。最初，人们确立了脑在思维(广义含义)和行为中的作用。随后确定的问题是，思维产生于脑质还是脑质中的脑孔(脑室)。接着，从脑的功能和解剖特征上对脑进行了粗略划分，如划分了小脑和脑皮质层。进而又

发现了脑皮质层自身各个区域的不同子功能，以及神经系统的基本细胞要素。最后确立的是这些细胞要素的结构和功能。我们就以这样的顺序分三步介绍：

(i)脑质与脑孔(脑室) 在第一个千年早期(公元100—400年左右)，理论家已经认识到脑是思维的器官这一重要原理，但却认为发生在脑内的那些孔道(脑室)之中。基于这种观点，“生命精气”储存在脑室之中，当需要时则通过中空的神经作用肌肉执行，以及获得感知(他们并没有说明这个过程如何进行)。这种观点以这种或那种形式持续了大约1500年(正确理论的出现着实不易!)。甚至笛卡尔也持有一种变更的脑室理论——松果体悬挂在中脑室，同时接受心灵和身体的作用。到了18世纪，思维才被完全确立为由脑质发生，但发生在脑的什么部位呢？又是如何发生的呢？

(ii)在什么部位发生？ 对于这一问题有两种回答，直到今天它们各自依然产生着影响。定位论起始于高尔(Franz Joseph Gall，1758—1828)，他提出这种观点的原因部分地出于其杰出的神经解剖学工作，部分地出于其哄骗性的骨相学说。高尔是最先用解剖学案例证明不同的认知功能大体上与脑皮层粗略的不同解剖部位相关，但是他认为可以通过这些不同部位的特征如测量脑颅突起部位的大小，推断个体具体的人格特征，如本书图3.5所示，这就走得太远了。

布洛卡(Paul Broca，1842—1880)和维尔尼克(Carl Wernicke，1848—1905)对于失语症(患者因脑损伤而丧失语言能力)历史性的研究，为脑功能定位论作出了重要贡献。随后，神经解剖学家如布罗德曼(Korminian Brodmann，1868—1918)等提出了更为详细的局部脑皮层与其功能的对应关系，如本书图3.5所示。

有很多理由认为，当代定位论者继承的这一传统可以追溯到两千年前的脑室定位论。“整体论”是与定位论相对立的另一种观点，代表性的人物有与布洛卡同时期的弗楼伦(Marie-Jean-Pierre Flourens，1794—1867)以及较近的拉什利(Carl Spencer Lashley，1890—1958)，但在生物学理论中从未占有主导地位。

(iii)如何发生？ 关于心灵机制的最大争论，主要体现在以发明水银染色技术而闻名的高尔基(Camillio Golgi，1843—1926)的“神经网络理论”与卡哈尔(Santiago Ramony Cajal，1852—1934)的“神经元理论”之间的争论。

争论的焦点是：神经系统与血液和氧化的循环系统一样，是一种完全的连续网络(高尔基：神经网络)，还是类似于人的骨骼带有间隙和连接点(卡哈尔：神经元)？显然，当代人们的使用习惯(神经元)已充分说明神经元理

论最终是正确的，尽管直到临近20世纪时因谢灵顿（Charles Sherrington，1857—1952）对于突触（卡哈尔的“节点”）的研究才取得决定性的证据。到了1950年前后，有关神经元的总体特征大体上都已经确定。

Ⅱ. CTM的现状：从图灵到2000年

图灵机　在1936/7年，图灵发表了《论可计算数及其在判定难题中的应用》一文，尽管标题有些模糊，但这也许是计算机科学中最为重要的一篇文献。在文中，图灵精确地论证了“(自动)计算机器”（为了纪念图灵，我们现在称之为“图灵机”）原理，并且进一步揭示了“通用图灵机”的概念——它可以做任何(其他)图灵机所能做的计算。随后，众多数学家和计算机科学家设计了在现实中能够执行更多任务的形式机器，但都被证明与图灵机是等价的。如果问(通用)图灵机究竟具有怎样的计算能力，可以粗略地认为：它能够计算任何能够计算的所有计算。

图灵测试　1950年，图灵发表了另一篇论文，题为《计算机器与智能》，对认知科学和CTM都产生了重要影响。在该文中，图灵探讨了机器具有智能的可能性，并提出了著名的“图灵测试”（他称之为“模拟游戏”）——粗略地讲，就是当人与具有某种智能层次的机器，如计算机，经过一段时间交谈后，将不能区分是与机器，还是与另外一个人在交谈。

A. 两个主题的融合(Ⅰ)：一篇作为转折点的文献

1943年，麦卡洛克（Warren McCulloch，1898—1969）和皮茨（Walter Pitts，1923—1969）在MIT发表了著名的但有些艰涩（甚至以专业标准看也是如此）的题为《神经活动内在概念的逻辑演算》一文。该文的重要性体现在两个方面：一是使多个学科产生了相互联系；二是促进了随后CTM的发展。

(i)使哪些学科联系在了一起？　这些学科主要包括神经科学、计算机科学和心理学（幸运地还包含些许哲学）。

神经科学　他们首先总结了当时神经科学所取得的一些主要成果，并“理想化”了神经元的一些重要特征（如本书图4.1所示）：

(1)神经元激活是一种“全一或一无”的过程；

(2)在潜伏附加期内，有固定数目的突触被激活，以便能够在任何时刻对神经元予以激活；

(3)在神经系统内，最重要的延迟是突触延迟；

(4)在任何抑制性突触活动时刻，神经元绝不会产生兴奋；

(5)神经网络的结构不随时间发生变化。

这种“理想化”的神经元模型对应着二值(真、假)命题逻辑的逻辑环路或逻辑闸，如本书图 4.4 和图 4.5 所示。

计算机科学　随后证明，当这种环路系统辅以可无限扩展的存储机制，那么系统就可以具有(通用)图灵机的计算能力。

心理学　提出所有的心理学问题归根结底都可以还原于神经系统的二值、开一关逻辑。

描述层次　该文还尝试对“信息加工系统”用三个不同描述层次对信息进行了统一和系统化。

(ii)产生的影响　麦卡洛克和皮茨的这篇文献，最终对认知科学的诞生以及进一步提出 CTM 产生了重要影响，是在冯·诺依曼的 EDVAC 报告中唯一引用的文献——EDVAC 报告实质上是所有随后数字计算机的蓝图，也是罗森布拉特(Frank Rosenblatt)感知器研究的基础——最终发展为联结主义研究，稍后我们再回到这个问题。

B. 两个主题的融合(Ⅱ):神经一逻辑

“神经一逻辑”主要研究类神经元网络的计算特征。

感知器及其缺陷　1958 年，罗伯森拉特发表论文讨论感知器问题，提出感知器是这样一种装置，“一旦被描述就产生感知”。感知器基本上是由麦卡洛克和皮茨神经元构成的一种网络，但除了具有神经元的上述五个特征外，网络还能够进行学习。一个简单的感知器包含三个部分：单元感觉输入层、单元联结层和单元反应输出层，如本书图 4.6 所示。

经过学习，感知器大致上能够与人类类似地识别出不同的刺激群，如男性面孔与女性面孔。但是在 1969 年，明斯基(Marvin Minsky)和帕佩尔特(Seymour Papert)证明，这种感知器在原则上被训练学习某些基本的判别(例如互斥问题，或者说要么 P，要么 Q，但不可两者同时)，从而导致了感知器研究的冷落，研究者们的兴趣点开始转向由明斯基和帕佩尔特提出的更具普遍意义的数字范式。

C. 两个主题的融合(Ⅲ):数字

图灵机　上文已有所提及，图灵为证明可计算和非可计算论题，在 1936/7 年提出了图灵机模型，是一种(相对)简单的抽象形式“机器”。可看作主要由两个部分构成：带有程序的磁头和一条无限长的磁带，磁带是机器

的“记忆”，能够依据程序指令左右移动，读写及删除字符，如本书图 6.1 和图 6.2 所示。

图灵机具有一种特殊的操作循环：（从磁带中）**读取**、（在磁带上）**印写**、**移动磁头**（每次一个方格）、（按照磁头中的程序）**进入下一种状态**。正是基于这样一种机器，图灵论证了他的著名论题。

冯·诺依曼机 1945 年 6 月，约翰·冯·诺依曼（John von Neumann，1903—1957，约翰并不是他最初的名字）完成了最初 EDVAC 设计报告草稿（这份报告草稿提交时尚不完整，有很多标识参考文献的空格，冯·诺依曼期望之后补充）。随后，EDVAC 计算机在宾夕法尼亚大学摩尔电气工程学院迅速流传开来，事实上成为以后所有计算机生产线的蓝图。如他本人所描述的，冯·诺依曼机具有如下特征：

（1）它是全自动的；

（2）具有存储数据和指令的记忆功能；

（3）能够储存程序；

（4）具有执行指令的控制机制；

（5）具有逻辑和算法的“器官”；

（6）具有输入和输出设置。

描述层次 值得注意的是，图灵机和冯·诺依曼机都可以看作是一种实际的“信息加工系统”，能够在两个层次上进行描述：物理环路（硬件）和运行的程序（软件）。

D. 两个主题的融合（Ⅳ）：联结主义

只有一层可训练的麦卡洛克和皮茨“神经元”的简单感知器，被证明不能学习识别某些对象类别，因而人们停止了对感知器的研究。包含多层“神经元”的感知器或许能够学习识别那些类别，但在一段时期内没有人知道如何训练它们。大约在 1985 年，众多研究者开始关注所谓的“联结主义”或者“并行分布式加工”的研究，此时业已证明能够训练多层神经网络识别那些疑难的类别。

三层前馈网络 随后，联结主义者提出了很多不同的网络组织或者“结构”，但我们注意到有一种是应用最广的，那就是具有三个联结层次的前馈网络，它应用于著名的 NETtalk 演示项目中的示例，如本书图 10.2 所示。

在 NETtalk 机低端的“输入层”给机器呈现书面文本，接着这些刺激产生的激活便传递到中间层或“隐单元层”，最后在机器顶端的“输出层”开始“朗读”这个文本。也就是通过调节连接权值从而使文本层与发音层相匹

配,训练 NETtalk 学习基本英文的发音。如果这一过程是靠手动完成的,那么我们自然地会称之为“程序运行”的步骤。三层前馈机器的这些特征表明,它们同样具有图灵机的计算能力。

描述层次 我们注意到,联结网络与感知器一样,具有类似的生物硬件,当给予与人类相似的心理任务时,它们通过调整各层之间的联结权值而能够很好地学习完成任务。

E. 心智计算理论(结果)

上述这些内容反映了什么“心智理论”呢?对这个问题的回答有助于进行历史的回顾,并了解我们现在所处的历史位置。

(i)心智表征理论(RTM) 至少从英国经验主义那里就具有一种“外在(out there)”心灵的观念,RTM(representational theory of mind)主要包含两个中心假设:

(RTM)

(1)认知的心理状态是表征的——我们通过表征世界而思维世界;

(2)思维过程最终都是这些心理表征的转换或“操作”。

对于经验主义而言,这些表征通常称之为“观念”,主要被认为是与它们相似的关于某事物的图像。我们在前面了解到,联想主义者认为思维主要是通过联想原则控制的“观念”连续过程。后来人们认为,表征的“相似性”理论与单纯的思维“联想”理论对于人类所有认知和行为的解释是不充分的。例如,我们已经了解到米勒、加兰特尔与普利布拉姆利如何利用复杂的 TOTE 构建他们的心智理论。

(ii)心智计算理论(CTM) 如果 TOTE 单元可以被看作是具体的程序示例,那么像脑一样操作 TOTE 单元的硬件便被看作是运行程序的硬件。因此,出现了著名的类比:**心灵之于脑正如软件之于硬件**。也有人不仅仅把这看作是一种类比,还把它当作一种科学宣言:心灵就是脑中运行的程序,并有这样的推论:想要得到人工心灵,唯一需要做的事就是找到合适的运行程序的硬件。既然现代数字计算机等价于通用图灵机,在原则上可运行执行任何计算的程序,这就意味着如果我们获得正确的心灵算法,那么就可以在数字计算机上编程使之具有心灵。我们如何知道它具有心灵呢?回答是如果能够通过适当的图灵测试则就证明计算机具有智能。这样我们就有了如下几种观念,塞尔(John R. Searle)曾称之为“强人工智能”:

(1)编程了的数字计算机,如果能够通过图灵测试,那么机器就具

有了智能、认知、心灵等；

(2)人脑基本上就是麦卡洛克一皮茨逻辑环路网络(十无限容量的存储=通用图灵机)；

(3)因此，可以通过找到我们脑内运行的程序，探究人类的智能、认知、心灵等。

请注意，这组命题并不是连贯的，前两个也许是正确的，但第三个是错的。世界上，智能(认知、心灵等)的来源可能有两种，一种出自程序，另一种不是，而人脑恰恰可能属于非程序的那种智能。因此，有人说图灵是 CTM 之父，这是错误的。图灵想要得到的是**机器的智能**，而 CTM 追求的是**智能的机器**。但 CTM 吸引人的地方是，很难想象还有另一种智能、认知、心灵等的来源。其他的任何一种理论看起来都具有神秘色彩，如福多(Jerry Fordor)所言："这是城镇里的唯一游戏。"

那么，用计算进行重新装扮 RTM，就得到了 CTM：

(CTM)

(1)认知状态是具有内容的计算心理表征的计算关系。

(2)认知过程(认知状态的改变)是具有内容的计算心理表征的计算操作。

我们由此就获得了 CTM 的工作状态。

(iii)数字 CTM 与联结 CTM 注意到，在介绍 CTM 时我们只提到了"计算"状态、表征、程序等。尽管 CTM 从经验上产生自传统数字计算机，但是并不保证一定要求如此——联结主义机器同样与比如冯·诺依曼机一样，是一种计算机。一种理论，如果要求计算结构和表征都是数字的，是一种有关数字的 CTM 理论，称之为心智数字计算理论(DCTM)；同样的，如果要求计算结构和表征都是联结主义的，是一种有关联结主义的 CTM 理论，则称之为心智联结计算理论(CCTM)。尽管当下人们对哪一种理论是正确的或者是更正确的存在很大争论，但它们都属于 CTM 的一种类别，而 CTM 又属于 RTM 的一种类别。

Ⅲ. CTM 的未来

共性难题

两种 CTM 的理论似乎都面对一些严重困难——有一些可以说是所有

心智理论都共有的难题。

首先是意识难题。如何能够使用计算的术语解释我们全部的意识经验——无论是哪一种机器，我们如何对其“编程”？

其次是表征或内容难题。如何用计算的术语解释某一符号关涉其事实所关涉的内容？典型的情形是，程序员们说如果 CTM 是正确的就行得通，但并不是所有人都对此满意。

数字 CTM 难题

软件难题 如何解释数字模型在众多层面上与人类心理相似性上的失败？我们能够列出很长的清单，表明普通人的心理功能对于数字计算机来说是完全陌生的。例如，(i)当有人受到微小损伤时，并不会“崩溃”，他会说“只是稍微有些不便”，以及(ii)我们似乎并不是通过“地址”检索记忆中的信息，而是通过内容进行检索——是关于什么的信息。

硬件难题 如何解释脑与计算机硬件(如芯片)在结构和功能上存在的巨大差异？如何想象一些标准程序以每秒百万次的速度在脑内运行？这只比迄今速度最快的数字计算机慢一点点。

联结主义 CTM 难题

软件难题 如何解释联结主义模型在众多层面上与人类心理相似性上的失败？例如，我们能够处理那些复杂的并且即时的加工任务，如语言加工、计划和演绎推理。

硬件难题 尽管联结主义模型是受“生物学”的启发而提出的，但它们仅仅粗略地模拟神经系统的结构和功能——如何使之更加精确呢？例如，训练三层网络使用的路径，并找不到与之对应的生物学知识。

展　望

数字和联结主义 CTM 的优缺点现在都已经显现得很清楚了。如果能够找到一种方式，将两者作为整体充分地结合在一起，那么 CTM 的前景无疑是非常乐观的：CCTM 应用于感知、记忆以及其他“直觉性”任务；DCTM 应用于推理、计划和其他“高级心理”任务。不过，目前还没有人知道究竟如何做才能实现。

R. M. 哈尼什

2010 年 4 月于美国亚利桑那大学

前　言

认知科学是研究认知的科学，由心理学、计算机科学、神经科学、语言学 xv 以及哲学等不同领域的学科所构成。促使这些学科结合在一起的是这样一种理念，即脑（神经科学）是一种计算装置（计算机科学），而认知（心理学）研究其软件——运行在脑中的程序。本书是关于认知科学**基础**的**历史**导论。这里的每个概念都很重要。说它是历史的，是因为我们将追溯一些关键概念的历史——基本上从19世纪（偶尔会回到亚里士多德）到现在。说它是基础的，是因为本书主要关注当代认知科学中研究心－身关系的一般（计算的）框架，并未涉及一些特殊的竞争理论及其包含的实验结果细节。同时，由于认知科学是一门交叉学科，我们将较为详细地介绍哲学、心理学、神经科学和计算机科学对这门学科的一些贡献。

导论：什么是认知科学，描述了关于这门学科的广义和狭义两种理解，并提供一种整合的观点。

第一部分：历史背景，追溯哲学、心理学以及神经科学对认知科学的一些贡献，直到大约1950年数字计算机的出现。

第二部分：心智的数字计算理论，主要关注一种特殊的人工智能程序——SHRDLU，并讨论一些数字结构和标准的知识表征方式。根据这些计算知识考察它们所蕴含的心智理论，以及关于SHRDLU的一些最重要的问题。

第三部分：心智的联结计算理论，进一步关注联结主义机器的心智模型。我们将考察两种联结主义计算解释程序——Jets & Sharks网络模型以及NETtalk模型。我们还论及联结主义结构、学习程序和表征程式等相关 xvi 内容。在讨论这些问题之后，将转向探讨它们可能蕴含的心智理论以及一些重要的相关问题。

结语：认知科学的计算，或者究竟什么是计算机？在分析计算机和计算两个重要概念后，我们试图对它们进行整合。结果表明，整合后的特征同样适用于数字模型和联结主义模型。

致 谢

本书是在我多年来在认知科学领域的教学讲义和演讲文稿的基础上完成的。首先应当感谢的是，许多使用过这些相关材料进行学习的学生。倘若没有他们的鼓励和支持，本书或许不是这个样子，也许根本就不会有这本书了。在写作过程中，我得到了下列优秀的教学和科研助手的帮助：彼得·格雷姆(Peter Graham)、拉莱·奎恩(Laleh Quinn)、邦莱·薮科(Bongrae Seok)、斯科特·亨德里克斯(Scott Hendricks)，以及布拉德·汤普逊(Brad Thompson)，他为本书编写了索引。最需要感谢的，是将我引入当时尚未称作认知科学的这一研究领域，并对我的学术观点产生重要影响的老师，他们是(以年庚为序)：加州大学伯克利分校(Berkeley)的伯特·德莱夫斯(Bert Dreyfus)和约翰·塞尔(John Searle)，麻省理工学院(MIT)的杰瑞·福多(Jerry Fodor)和内德·布洛克(Ned Block)。把从他们那里所学到的东西融会贯通，对于我来说是一种挑战。亚利桑那大学(University of Arizona)以及其他大学和研究机构的许多同事审阅了本书的部分章节，并提出了非常有益的反馈意见，他们是(按照姓氏字母先后顺序)：肯特·巴赫(Kent Bach)、汤姆·贝弗(Tom Bever)、内德·布洛克、安德雷·伯茨(András Bocz)、迪克·卡特(Dick Carter)、戴维·查尔默斯(Dave Chalmers)、肯恩·弗斯特(Ken Forster)、梅里尔·加勒特(Merrill Garrett)、阿尔文·戈德曼(Alvin Goldman)、比尔·伊特尔森(Bill Ittelson)、汤姆·拉尔森(Tom Larson)、肖翰·拉维恩(Shaughan Lavine)、克里斯·马洛尼(Chris Maloney)、林恩·纳代尔(Lynn Nadel)和里奇·泽梅尔(Rich Zemel)。詹森·巴尔克(Jason Barker)帮助处理了一些插图，艾格斯·帕斯特(Ágnes Pásztor)帮助获得了本书封面图片的印刷使用授权。亚利桑那大学非常适合从事像本书

xvii

所关注的这种跨学科研究工作，特向认知科学研究中心主任加勒特致以谢意，感谢他提供的这种环境。感谢萨巴·普莱赫（Csaba Pléh）邀请我到布达佩斯大学（ELTE）讨论书稿，同时向洛克菲勒基金（Rockefeller Foundation）表示感谢，该基金资助我到贝拉吉欧研究与会议中心（Bellagio Study and Conference Center）进行了为期一个月的访问，为本书的最后完成提供了一次难得的机会。最后，我的夫人塞拉·帕斯特（Scilla Pásztor）对我的支持远远不是感谢所能表达的——将此书献给她。

关于本书引文的说明：所有引文均出自参考文献和推荐读物，但为方便阅读，书中只对一些（有趣的、有争议的、重要的和令人惊奇的）引文注明了出处。

目　录

第二部分　心智的数字计算理论

第三部分 心智的联结计算理论

插图目录

导论：什么是认知科学？

认知科学(cognitive science)，作为建制化的研究领域，是20世纪70年代中期之后由于既有的各种认知研究领域之间存在鸿沟而产生的。虽然认知是当时人们进行研究和关注的核心，但是还没有关于认知的科学，没有关于思维领域的系统研究：知识如何获取？又如何在心智中进行表征？知识如何应用到思维和行为中？如何提高知识获取和利用的功效，以及如何克服这些问题可能出现的缺陷？正如计算机科学始于对电气工程和应用数学进行整合的特殊建制程式，"认知科学"自觉地成为一门具体学科，也同样包含了许多不同领域研究者所关注的问题和运用的方法，诸如心理学、神经科学、计算机科学、语言学、人类学和哲学等学科的相互交叉。70年代后期和80年代早期，这些研究领域的研究者开始汇聚起来，共同致力于认知问题的研究，于是"认知科学"作为一门学科建立起来。在此过程中，有如下三个里程碑式的事件： 1

1977 《认知科学》杂志创刊

1978 斯隆基金会报告：认知科学的学科状态

1979 认知科学学会：第一次会议(La Jolla, CA)

从这时起，将认知科学当作一个"正式的"研究领域似乎是合理的。也很显然，认知科学将从与其他学科的系统联系中获益，因为其他学科的学者也有兴趣从自己的视角提出对认知的理解。现在对"认知科学"主要有两种解释——用较为准确的术语表达，我们可以分别将其称为**广义理解**和**狭义理解**。既然我们将采取把这两种解释进行整合的进路，那么我们就应当分别阐述它们各自的含义，以及它们各自的优势和缺点。

Ⅰ.1 广义的理解

2 粗略地讲,"认知科学"的广义理解就是研究认知的科学。它基本上是这样一个学科,包括一个研究**领域**以及研究该领域的一套**规程**。对于这种理解,诺曼(Norman, 1981:1)的观点颇有代表性,他认为:"认知科学是一门新兴的学科,它产生于从不同观点对认知进行研究的汇流。认知科学的关键是寻求对认知的理解,现实存在的或抽象的认知、人的或机器的认知都是其所关注的对象。它的目标是探寻智能和认知行为的原理。认知科学可望获得对人的心灵、教学与学习以及心智能力的理解,并且能够通过重要的和积极的方式发展智能装置以增强人类能力等问题的理解"。考虑到认知所涉及的领域,其核心学科(哲学后面述及)包括:(1)认知心理学;(2)认知神经科学;(3)计算机科学;(4)语言学;(5)人类学。

认知心理学 为我们提供了许多关于各种认知能力的精细理论,同时给出了评价这些理论的实验范例(参见:Bower and Clapper, 1989)。

认知神经科学 为我们提供了对于支撑认知能力的神经系统的部分分析。它告诉我们具有特殊神经特征的系统如何具有它们所附有的认知能力。

计算机科学 为我们提供了对于复杂的(智能的?)能力,如何在物理系统中实现这一问题最全面的理解——软件如何与硬件相联系。对于计算机的各种结构、算法和数据结构的研究,为理解生命体系是如何组织起来的这一问题提供了潜在的理论见解。

语言学 对于认知科学来说,从历史上看它的作用是直接的,但是从它的主题上看,它的作用则是间接的。在20世纪50年代后期和60年代早期,语言学家如乔姆斯基(Chomsky)对行为主义的批判,影响了一代心理学家(如乔治·米勒)和哲学家(如希拉里·普特南)。乔姆斯基的转换语法理论提出存在大量的认知结构,很多认知科学家也一直 3 在研究这些思想。语言加工是一项重要的认知能力,幸运的是,已经有专门的学科(语言学)在致力于研究语言的输入与输出。

人类学 在认知科学中起着独特的作用,它从跨文化的视角探究认知现象,提供了它独特的观点。

哲学 在认知科学中具有显著的作用。首先,哲学不是一门科学,所以也就没有"科学的"方法可以提供。但是,现在许多认知科学面对

的问题都是传统的哲学问题，如心一身问题、人格同一性问题、意识、心理表征以及理性等。因为这些问题非常基本并具有普遍意义，所以难以运用具体的科学方法解决，而那些具体的科学则更多地用于解决局部问题。哲学长期关注上述问题，虽然还没有解决它们，但是已经描绘了采取不同的进路解决这些问题可能会出现的一些结果。哲学的分析方法，对于促使在这些科学研究领域里获得精致的见解颇有助益。

这些相邻学科是如何联系在一起的呢？一种较普遍的观点认为是源于1978年斯隆学科状况报告提及的“认知六边形”（参见本章附录）。那么，**广义理解**又可以近似地看作，认知科学是**依据这六门学科的方法对认知展开的科学研究**。当然，相互褒扬并不会总占支配地位，合作也可能出现破裂，正如丹尼特（Dennett）戏谑地说道：

> 我们难道会去询问人工智能领域的人们，你们会浪费时间同神经科学家讨论吗？神经科学家并不谈论什么“信息处理”，而只关心在**哪里**发生，涉及了哪些神经递质以及那些琐碎的事实，但是他们却对高级认知功能的计算要求一无所知。我们难道会去询问神经科学家，你们会浪费时间进行人工智能的幻想吗？人工智能研究者只是在发明他们所需要的机器，并显示出对大脑认识的不可宽恕的无知。与此同时，认知心理学家的调和模型也受到指责，这个模型**既不**具有生物可能性，**也不**具有经验证实的计算能力；人类学家即使看到了这个模型，他们也不会关心；哲学家，正如我们所知，总是互相批判和诘难，为他们自己制造的困惑而焦虑，同时在他们的舞台上，不但没有数据，而且他们的理论也不具有经验上的可验证性（Dennett, 1995: 254-5）。

Ⅰ.2　狭义的理解

狭义理解的认知科学，是说它并不是一个研究领域，而只是一种**理论假设**（doctrine）。这种理论假设的基础是心智计算理论（computational theory of mind，CTM）——心智/脑是某种类型的计算机。引用1978年关于认知科学的斯隆报告：“认知科学所有分支学科共同拥有……一个一般研究目标：**探索心智的表征和计算能力，以及它们在脑中的结构和功能表征**”（参见：Sloan Report，1978:76）。一些研究者明确地赞同这种狭义理解，而反对广义的对认知科学的构想：　4

> 当说到认知科学时，我脑中会出现这样的想法……它包括认知心理学和人工智能的一些重要研究预设，在一定程度上还包括语言学和哲学。很多——也许是大多数——认知心理学家都明确地不赞同这种观点，尽管多数研究者从事着具有这种特征的研究工作。在这个问题上，很多人喜欢用“认知主义”来说明我的这种想法，把“认知科学”看作是一个中性术语。我认为，“认知科学”是一个理念附载术语（如同“社会生物学”），因为**如果没有那种理念，认知科学旗下进行的所有研究项目根本就不会形成一个专门的研究领域**。(Block，1983：521)

这种评论当然有其合理之处。并不是把一些研究项目聚拢在一起，就能形成一个研究领域。但是，作者所强调的如果没有那种理念就不会形成认知科学的研究领域，这一观点能否令人信服？我们会同意吗？也许同意，也许会不同意。我们知道，**理论假设的变更针对的是它的研究主题**，而并非是指“研究项目”。此外，在人类学，或者历史学，或者哲学，甚至语言学这些领域，都不可能只有一个研究项目就可以构成——广义理解和狭义理解，认知科学究竟选择哪一种更好呢？

Ⅰ.3 认知：广义与狭义

尽管广义理解的认知科学中的各门学科都提供了与“科学”相关的概念（这些学科所做的事，就是“科学”），但我们仍不知道哪部分心理活动属于**认知**。作为心理现象的“认知”究竟是什么呢？像许多其他重要概念一样，“认知”似乎拥有许多显然的例子，但是缺少一种普遍性的定义。对于什么是认知，其中一种界定方法与对认知科学的广义理解相似——只是罗列出有关认知现象的一些清楚的具体问题，然后宣称，所谓认知就是研究这些问题。正如很多认知心理学教科书所说的那样，通常认为认知领域应该包括如下子领域：注意、记忆、学习、推理、问题求解，以及动机理论和行为理论的某些方面。还可以将感知和语言加工方面的一些问题，列入这个清单。所以，我们可以说，**广义认知**就是对注意、记忆、学习、推理、问题求解，以及动机理论、行为理论、感知和语言加工的某些方面所做的研究。但是，这些问题具有哪些共同之处呢？怎样知道何时我们应该为这份清单增添新的内容呢？

上面罗列的这些问题似乎都具有这样一个特征，就是它们都包含着某种形式的心理表征的心理“操作”。我们将其定义为**狭义认知**：认知就是心

理表征所做的心理"**操作**"(生成、转换和删除)。例如,感知的内容典型地表征了引起感知觉的对象。记忆(真实的)建立在某种感知基础上,还涉及语言,所以记忆必然与感知本身表征的对象、事件和情境有关。其他的认知功能也具有相同的表征特征。某人通常对"某事"进行推理或者计划怎样"做某事",这两者都需要对世界的本来面貌,或者世界的可能面貌,或者世界将要出现的面貌等进行表征。

广义认知与狭义认知之间具有什么关系呢?我们的**基础假设**是,上述广义理解的认知现象都包含着狭义理解的上述关于认知的描述。当然,这可能会被证明是错误的(行为主义就对此持有异议),但认知科学总是要对这些现象进行研究。

Ⅰ.4 计算:广义与狭义

广义理解的计算,仅指计算机之所做。这样当然就表示把所有具体的计算特征,都归结于计算机之所能做(计算机当然还可以做很多其他事情,比如散热),并未论及计算机究竟是什么(你的数字手表是计算机吗?)。一种普遍的观点是,计算机主要用于输入、储存和输出"信息"——它们是一种"信息加工"装置。但什么是信息加工?这些装置具有什么共同之处呢(即使仅是一种家族相似)?一种很有影响力的**狭义**计算机概念是马尔(Marr)提出的"三层次"假设:信息加工装置就是运算符号操作器,这种操作器可以描述为三个不同的重要层次,它们需要分别回答三种不同的重要问题:系统能够解决什么问题,使用哪种算法规则,以及如何在物理世界中(硅、神经组织以及其他)实现? 6

Ⅰ.5 认知科学的基础概念

狭义认知科学、**认知**和**计算**三者,可以通过如下方式结合在一起:如果认知是心理表征所做的心理操作,并且如果这些表征是符号的,这种操作是自动的,那么认知就是指一系列计算,这就是狭义认知科学。既然认知科学是一门跨学科研究认知的科学,很明显认知科学的广义和狭义理解并不等同,但是认知也很有可能不是计算。采用狭义理解也许是错误的,而如果这种理论被证明是错误的,那么就会把自己置于一种没有主题的境地。有人

同时采用这**两种方式**理解认知科学的研究领域。例如，加德纳(Gardner，1985：6)列出认知科学具有如下四点核心特征：

1. 当谈及(人类)认知活动时，就需要提到独立于生物学、神经学和社会学的心理表征；
2. 理解人类心智的关键是计算机；计算机是探索人类心智功能最为可行的模型；
3. 认知科学并不强调情感、历史和文化背景等方面的因素；
4. 认知科学是一种交叉学科。

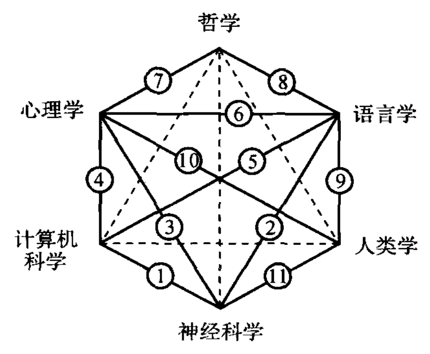
这里，我们看到第 2 条所述是对认知科学的狭义理解，即探索认知计算的学科；第 4 条是对认知科学的广义理解，即主张对认知进行跨学科研究。根据本书提出的“认知科学的操作概念”，我们对认知科学的广义和狭义两种理解的关系总结为：认知科学的**研究对象**是认知，认知科学的**方法**是组成认知科学各门学科自身的方法(见认知六边形)，该领域的**核心假设**为心理状态和过程可计算。按照这种理解方式，我们凸显了认知可计算的观念，同时确保如果这种观念被证明是错误的，认知科学也不会自我塌陷。

7

附录　1978 年斯隆报告(摘录)

认知科学研究智能实体与外界环境互动的原理。理所当然，这种研究必将超越学科的界限，将神经科学、计算机科学、心理学、哲学、语言学和人类学等学科的研究者们所做的工作整合在一起。对这些学科的熟悉就为人们探索认知科学的研究状态，提供了适用路线图。

在过去的十多年间，各个子领域之间密切的联结方式，已经清晰地表明，对认知进行研究已经作为一门独立的学科出现。组成这门学科的各个学科及其相互间的关联，可概括为图Ⅰ.1。图中所列六个领域中的每一个领域，都通过跨学科动态网络同其他领域连接在一起，其中因交叉产生的一些问题是自古以来人们就一直关注的，而另一些还没有在主流学术界成为焦点，但正在逐渐被我们所熟悉和重视。



认知科学的子域:(1)控制论;(2)神经语言学;(3)神经心理学;(4)认知程序模拟;(5)计算语言学;(6)心理语言学;(7)心理哲学;(8)语言哲学;(9)人类语言学;(10)认知人类学;(11)大脑的进化

图Ⅰ.1　认知六边形(来源:Pylyshyn, 1983: 76)

这些领域的每一个组成部分,通过跨学科动态网络与其他两个或多个领域相连接。图中每一个连线中的圆圈标记,都表示已经有明确界限的研究领域,该研究领域与连线两端的学科涉及在知识或物理工具上的关联。因此,控制论运用的思想,是由计算机科学家模拟人脑的功能而发展起来的,而人脑的功能则是由神经科学家阐明的。类似地,心理语言学连接心理学和语言学两个领域,关注于主导语言的习得、产生和理解所需要的心理机制和心理操作;认知程序模拟连接计算科学和心理学,试图将思维和问题求解明确进行公式化描述。其他的由两个学科的连接而形成的领域,其情形都相类似。

8

图Ⅰ.1中有11条实线,每一条都代表一门已经明确定义,并且建立起来的专业化的跨学科研究领域,在传统的学术教学部门中都可以找到一个或者多个这种研究领域。图中四条虚线标示的学科之间的联系,展现了一系列新的问题,某些已经为人所熟悉并变得重要,但是还没被学界正式认可为专门的研究焦点。

我们还可以同时把六个主要学科中的三个或者更多组合结集。比如,将哲学、心理学和语言学三个学科组合起来形成的学科组,就代表了其关注的是认知任务中的语言及其应用,这样一个较为传统的研究领域。每一个这样的学科组都代表某一根基牢固,且正逐步活跃的研究领域,研究者都要受到过两门或者多门学科领域的训练。本文主要关注的是,阐明我们的观点,即在此所展现的具有相互影响的各门学科所形成的网络应当被当作一个整体,并冠以认知科学的名称。虽然这个整体尚未被成功整合,但这种整合却是所有相关研究组群所朝向的目标。

的确,认知科学的所有分支学科共同享有的,也就是使得认知科学领域得以存在的主要原因,在于它拥有一个共同的研究目标:**探索心智的表征和**

计算能力，以及它们在脑中的结构和功能表征。认知科学已经和正在被上述研究领域和子领域的研究者们所实践。这些研究者们已经接受了这样的挑战，那就是如何对认知系统进行充分的理论阐述，并且对这些理论的预言进行经验证实。

(Sloan Report：75-6)

【思考题】

认知科学的产生是为了满足什么可感知的需求？
什么是“广义的”认知科学？
认知科学包含哪些主要学科？
这些学科都对认知科学具有什么贡献？
9 广义的认知科学的主要问题什么？
什么是“狭义的”认知科学？
狭义认知科学的主要问题是什么？
什么是广义的认知？
什么是狭义的认知？
什么是广义的计算？
什么是狭义的计算？
狭义的认知和计算是如何构成狭义认知科学的？
什么是认知科学的操作概念？

【推荐读物】

认知科学的特性

参见例如 Pylyshyn (1983)及其评论，特别是 Newell (1983)。关于认知科学家的特征，参见 Baumgartner and Payr (1995)，该书颇具启发性(引人入胜)。

认知科学的介绍

有大量的、优秀的介绍认知科学的读物，涉及前面评论提到的基于广义和狭义两种理解的认知科学。出于好奇，可以仅翻阅 Wilson and Keil (1999)这本引人入胜的导论性著作。Dunlop and Fetzer (1993)则是一本对认知科学中最重要的概念给出定义的有益的、简洁的读本。

广义的认知科学

Stillings et al. (1995)是第一本认知科学教科书，它纵览了多个学科及其对认知科学作出的贡献。该书写作清晰，由多名作者撰写，但是却具有单独作者所著文字具

有的优点。多卷本著作 Osherson et al.（1990，1995）涉及认知科学中的多个主题，该书每一章都是由该领域的著名权威学者所单独撰写的。

狭义的认知科学

Johnson-Laird（1988）是基于将心智概念理解为一种计算装置的理念而写作的最早教科书之一。较为近期且基于同样的理念而写作，但所牵涉的题材与我们较为相近的教科书为 von Eckhardt（1993）。该书第 1 章和第 2 章包含有关认知科学的本性的大量讨论。同样地，Crane（1995）以一种可读的和生动的方式包含了一些我们的非历史题材——表征的特性是该书关注的焦点。Thagard（1996）作者围绕认知科学中运用的多种表征形式撰写了该书的导论。Dawson（1998）则围绕马尔的“三层假设”而展开写作（参见上文）。 10

认知科学的历史

迄今为止，最著名且颇具可读性的认知科学史著作为 Gardner（1985）。Flanagan（1991）也包含一些有趣的历史章节，既有我们讨论到的人物（如威廉·詹姆斯）和活动（如行为主义），也有我们未涉及的内容（如弗洛伊德、格式塔理论）。Bara（1995）的第 1 部分也回顾了认知科学的历史，还有 Bechtel and Graham（1998）的第 I 部分也是如此，这是一部优秀的认知科学简史著作。

文集

Haugeland（1997）是一本极佳的论文集，收集了一些有影响的、原创性的研究论文，而 Posner（1989）则是一本汇集了认知科学基础理论研究论文的优秀文集。Collins and Smith（1988）主要关注心理学和人工智能。Garfield（1990）选编自多种资源，并突出了哲学议题。Goldman（1993c）是一本选题广泛并强调认知科学对哲学的影响的论文集。Thagard（1998）为 Thagard（1996）提供了一本有用的指南，Cummins and Cummins（1999），以及 Lepore and Pylyshyn（1999）则是与此相关的最新著作。

相关学科

实际上，所有的**认知心理学**著作都与认知科学有关，因此我们将不对其进行考察。我们会提及 Barsalou（1992），该书试图将作为概述的 Baars（1986）与从认知科学创立者之一的视角而进行考察的 Hinst（1988）建立起联系。对**人工智能**的哲学探讨经常与认知科学的某些部分相互交叠。参见 Copeland（1993b）或者篇幅相对较短并较少涉及技术问题的 Moody（1993）。Boden（1990）选集收入了本书讨论到的许多篇文章。**心灵哲学**的探讨也与认知科学相互交叠。在文本方面，参见 Churchland（1988），Sterelny（1990），Kim（1996），Braddon-Mitchell and Jackson（1996），Goldberg and Pessin（1997），以及 Rey（1997）。文集方面，参见 Block（1981） and Lycan（1990）。Goldman（1993d）追溯了认知科学对一些哲学分支学科的影响，

Guttenplan (1994)是一本很好的包括许多与认知科学直接相关的论文的手册。在**神**
11 **经科学**方面,P. S. Churchland (1986)对其进行了详细考察,并研究了它与哲学的关系;Churchland and Sejnowsky (1992)试图整合神经科学与认知科学。Gazzaniga (1995)是关于认知神经科学的主题、问题和领域的重要文集。Squire and Kosslyn (1998)是一本篇幅较短的主要收录关注认知神经科学子域的近期文章的文集。在人类学的认知方面,可参见 D'Andrade (1989)。如果想快速了解当前认知科学的视野,可以阅读新近出版的《认知科学学会年度会议文集》。

第一部分

历史背景



引 论

心智的计算理论(computational theory of mind, CTM)最早以“数字”形式出现,后来又出现了“联结主义”的形式。本书第一部分的目的是全面考察CTM形成的某些历史影响因素。当然,并不是我们要讨论的所有领域都对CTM具有直接(甚或间接)的贡献,但是这些领域都与广义理解的认知科学相关。从这幅视野开阔的画面中,我们尝试描绘出,这些领域对CTM有何直接影响的一幅简图。我们将看到,这些领域对形成认知科学的影响,尤其是对后面将要详细介绍的心智计算理论的影响。CTM联结主义形式是从感知器(第4章)和联想主义(第1章)演化而来的。在介绍联想主义(第1章)之后,我们将转向联结主义(第12章)。从联想主义和詹姆斯那里,我们获知下列思想:(i)有意识的心智是表征(观念)的内省操作器(联想);(ii)有意识的心智具有两层解释,即内省的、主体的和心理学的(软件),以及客观的、神经学上的(硬件)。自行为主义产生以来,人们逐渐抛弃了单一内省研究,而转向实验室的实验研究,正如现在认知心理学所运用的实验研究方法。信息加工心理学认为,认知就是信息的加工,这一观点尤其得到米勒(Miller)、加兰特尔(Galanter)和普利布拉姆(Pribram)等人的赞同,并应用TOTE(Test-Operation-Test-Exit,测试一操作一测试一输出)单元解释人的行为是结构性的,其功能犹如计算机程序。 15

我们从生物学中获得了最重要的神经元原理:神经系统的存在,以及脑是由离散的单元(神经元、轴突、树突)通过突触连接而形成的网络所构成的。神经一逻辑的一般观点,特别是麦卡洛克(McCulloch)和皮茨(Pitts)认为,人脑是由单元以及单元环路的开/关构成,因此可以与逻辑命题相关联——人脑相当于图灵机的工作台,如果拥有了无限的记忆,那么就相当于一台普适图灵机。感知器则表明,参照人脑总体解剖结构组织起来的计算机硬件,通过训练能够以明显类似于人的方式识别事物的类别。

1 联想主义

1.1 引言:什么是联想主义?

16 “联想主义”认为,心灵是按照联想的原则组织起来的,至少在某种程度上如此。联想主义者并未指明“联想”原则源于何处,相反地,他们热衷于谈论具体的原则,将这些原则都称之为“联想”(“联想”一词由洛克最早提出,见下文)。但是,联想主义背后的**基本思想**似乎是:**在经验中“处在一起”的东西也将在思想中“处在一起”**。联想主义者通常都是**经验主义者**——他们坚持认为所有的知识均来自经验,既包括产生依赖于经验的知识,也包括仅仅由经验判断推理得出的知识。然而,这仅涉及一般公认的观点,事实上每一位特殊的经验主义者都持有与他人略有不同的观点。

1.2 一般经验联想主义

> 这些英国心理学家——他们究竟想得到什么呢?你总会发现他们……热衷于寻找确切有效和最直接的原因……那是以拥有智慧而自豪的人类最希望探寻的地方(例如,习惯的惯性,遗忘,以及如何无意识地或偶然地把各种观念结合到一起,或者有时表现出纯粹被动,无意识、反射、分子态和彻底的愚蠢行为)——究竟是什么使这些心理学家专注

于此？对于渺小的人类来说，这是不是一种神秘、恶毒、庸俗，也许还是自我毁灭的一种本能呢？

——尼采(Nietzsche，1887)

虽然经验主义者在心理结构和心理操作等概念的一些细节上存在意见分歧，但是，我们还是可以在一般的总体框架中理解它们，称之为“一般经验联想主义”。它虽然并不与任何一种经验主义完全符合，但是毕竟为我们提供了一幅经验主义的关于观念活动的合成图片(见图 1.1)。 17

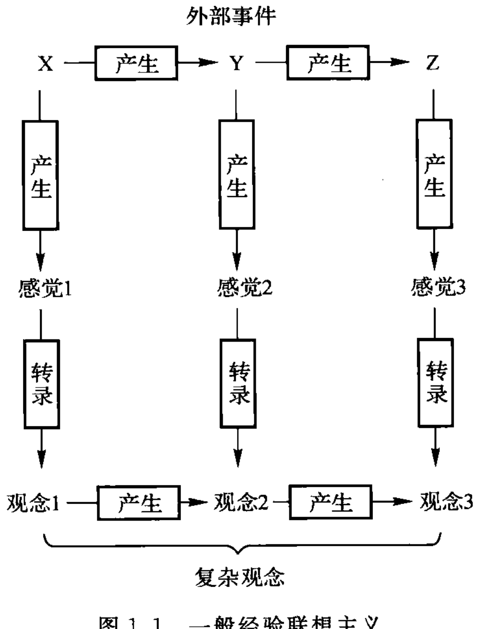


图 1.1　一般经验联想主义

基本原则

图 1.1 表明，一般联想主义至少包含三个基本原则：(i)观念间能够在心灵中相互联结，譬如(与上文相对照)，通过经验联结的东西也将在观念中联结；(ii)观念可以分解为若干“简单”观念，形成一个基础的“简单”观念储存盒，这个储存盒内的简单观念又可以构成一些更为复杂的观念；(iii)简单观念均产生自感觉。感觉(直觉材料)并非受联想原则支配，而是完全由人脑以外的事物引起(霍布斯、洛克、休谟：世界；贝克莱：上帝)。

作　　者	时间	邻近律	相似律	相对律	因果律
亚里士多德(Aristotle)	330 BC	X	X	X	
托马斯·霍布斯(Thomas Hobbes)	1651	X			
约翰·洛克(John Locke)	1700	X	X		
乔治·贝克莱(George Berkeley)	1733	X	X		
大卫·休谟(David Hume)	1739	X	X		(X)*
大卫·哈特利(David Hartley)	1749	X			
詹姆斯·密尔(James Mill)	1829	X			
约翰·斯图尔特·密尔(John Stuart Mill)	1843	X	X		
亚历山大·培因(Alexander Bain)	1855	X	X		
赫伯特·斯宾塞(Herbert Spencer)	1855	X	X		

* 其后归结为邻近律①。

图 1.2　联想的原则(来源:Marx and Hillix, 1963:106,图 8;经 MeGraw-Hill Companies 许可复制)

1.3　联想的种类

18　**完全**联想主义认为联想原则完全支配着心灵的操作,然而**混合**联想主义则认为除联想原则外,还存在非联想原则。联想主义认为事物存在多种维度的"处在一起(going together)",比如可以在空间上具有某种关系(如空间邻近),也可以在时间上具有某种关系(如时间相继)。还可以因一些较抽象的关系而联结,如因果、相似以及相反。下面是一些典型的联想原则:

1. *邻近律*:在空间或时间邻近的事物较易产生联想。
2. *相似律*:相似的事物会联想。
3. *相对律*:相对的事物被联想。
4. *因果律*:具有因果关系的事物会产生联想。

尽管联想主义者并未提供确切的论证来回答为什么上述联想原则不能无限扩展,但近年来却明显对操作联想主义原则取得了认同。

① 休谟认为大多数人都相信只要一件事物伴随着另一件事物而来,两件事物之间必然存在着一种关联,因果概念只不过是人们期待一件事物伴随另一件事物而来的想法。所以,本书作者认为休谟的因果关系实际上指的是邻近关系。——译者注

联想的过程

对于联想主义而言，有三种主要的联想过程。其中一种过程指，**哪一事物在时间中出现在另一事物之后**，例如回忆某事或者思维的时间顺序。另 19 一种过程则涉及**复合**，例如把若干简单的东西组织，结合为更复杂的事物。最后一种过程为**分解**，把复杂的事物分解，拆斥为简单的部分：

1. *相继*：联想原则支配着这一过程，如回忆某事，或思维顺序；
2. *复合*：复杂的事物通过(a)心理机制，(b)心理化学由简单的事物结合而成；
3. *分解*：复杂的事物分解为较简单的成分。

对于复合怎样产生，主要存在两种不同观点，一种以洛克为代表，认为产生于某种“心理机制”；另一种如米勒(见下文)，则认为产生于“心理化学”。

联想适用范围

最后不同的联想主义者都认为，联想原则涉及多种范围：记忆、观念、表象、思维，以及所有经提示的、使用过的和拒绝的东西[1]。对联想主义进行这种一般性的考察之后，我们再来看看这些思想如何由两位著名的联想主义者得出。

1.4 洛克与詹姆斯

最初对于联想主义的研究，很难区分出它们对哲学和心理学的各自贡献，因为哲学和心理学本就很难区分。尽管一些成果所关注的对象明显属于哲学(大卫·休谟)，另一些则明显是心理学所做(威廉·詹姆斯)，还有一些两者皆有(例如，大卫·哈特莱(David Hartley)，与洛克相似，既是一位医生，也是一位哲学家)。造成这种模糊不清的原因在于，一些哲学家写的东西像心理学家写的，而一些心理学家写的却像是哲学家所作。然而，似乎从哈特莱开始，联想主义作为经验主义认识论的一个分支，转变成心理学的基本原则。在霍布斯(Hobbes)之前，联想主义同心灵问题彼此不分。与此同时，笛卡尔提出了著名的天赋观念说(联想主义属于完全的神经学概念)。哲学的联想主义同英国经验主义(约 1700—1850)共同达到了其辉煌顶峰，

20 尽管对于联想主义最详尽的理论论述可能是詹姆斯·密尔(James Mill)的著作《人类心灵现象剖析》(1829),然而影响最广泛的哲学论证则可能是大卫·休谟的《人性论》(1739)了,但是当代认知科学家似乎更关注洛克和詹姆斯对传统联想主义的贡献。

约翰·洛克:《人类理智论》

约翰·洛克(John Locke,1632—1704)与波义耳(是其好友)和牛顿处于同一时代。他曾于牛津大学学习形而上学和逻辑,并在此有过一次用其话讲"丧失了理智"的恋爱,然恋爱未果,其理智似乎也并未消失。从此终生未婚,致力于撰写著名的关于知识理论的著作《人类理智论》(*An Essay on Human Understanding* (1700),用了 20 年写作完成)以及阐述他的政治理论(深刻影响了《独立宣言》)。

观念

洛克并不认同笛卡尔(见第 3 章)的天赋观念说,他说:"设想心智开始之时,如我常讲,似一张没有任何字迹的白纸一样没有任何观念。它是如何变得丰富多彩的呢? ……我的答案只有一个词——**经验**,基于此我们才得以获取知识,而它终究只能来源于自身"(《人类理智论》第 2 册第 1 章第 2 节)。心理内容(观念)既可来自于外部经验,如**感觉**,也可以来源于内部经验,如**内省**——心灵的自我操作。感觉和内省产生**简单**观念,通过诸如对其相似和差异的识别,或者抽象等心理操作手段,产生诸如本质、关系等**复杂**观念。

世界

感觉为我们提供了关于外部事物特征的各种观念,事物的特性有两种重要分类方法:

> *初性*(primary qualities)(例如硬度、延展性、形状、运动、静止、数量)是本质的、关键性的。它只有物质本身才具有,且完全独立于心灵认识。

21

> *次性*(secondary qualities)对于物质而言并不具本原性,是物质(通过初性的构造)在心灵认识中产生经验(诸如颜色、声音、味觉、嗅觉)的原因。

关于世界的观念

洛克说，人们会认为同时存在一个外部世界和一个内在自我，“通过感觉我们知道存在着实体的、广延的物质；通过内省知道我们拥有着像思想这种东西；经验则使我们确信物质与思想同时存在”(《人类理智论》第 23 章第 15 节)。但上述关于心灵内容的图式存有一个疑问，即我们怎样获得关于外部世界的知识(或者关于我们自己的)？洛克提出了一个表达这种窘境的隐喻，同时也指出了解决的方向，“可以这样理解心灵，它好比是一个封闭的暗室，留有一道缝隙，外物的映像或者关于事物的**观念**从缝隙进入；到达暗室中的映像或者观念如果能被储存，并且有秩序地存放以备迅速提取，那就如同人可以看见各种物体，并能理解有关物体的各种**观念**差不多了”(《人类理智论》第 11 章第 17 节)。这里主要有两种观念(后面将详述)：(1)一种可以类比于**画像**，是外界事物的“真实的映像”的感觉观念；(2)另一种是与事物相关的性质相连接的感觉观念。接着，洛克强调了这两种观念的区别：第一种观念**与事物初性相似**——是“这些性质的映像，这些外物原型的样式真实地存在于心灵之内”(《人类理智论》第 11 章第 17 节)。然而，“外物映像”的观念**并不与事物次性相似**——世界本身并不含有甜的味道或者蓝的颜色，它们仅仅是事物运动的扩展结果。然而，哪种连接对于表达事物相关性质的观念是必需的呢？他似乎认为是**因果**连接：“事物的种种现象状态作为一种标识，使我们能够认识和区分我们所接触的事物。**观念**具有同样的效果，并且具有能够真实地对事物进行区分的特征，而不论它们是否是事物的一些恒常的结果[次性]。另外，事物本身精确的映像[初性]：这一切都是由于观念与外界实物的真实构造**恒定契合**的缘故……这样，观念就可以由外物**恒常产生**”(《人类理智论》第 30 章第 2 节)。这就是我们怎样突破感觉的暗室通向外部世界的——因果和映像。在下面的章节我们还会更深入地对此进行谈论。

我们现在可以区分在洛克的理论中经常讨论的两个概念了：复合和连续。但并不确定在洛克那里联想原则所适用的确切范围(尽管我们通常假定为“观念”)，也不确定其联想原则的应用和涉及范围。洛克认为，存在三种普遍的心理操作：(1)结合；(2)使观念相互邻近，但并不结合(关系观念)；(3)抽象(普遍概念)。第一种关系与联想最为接近。 22

结合和复杂观念

对于发生在心灵内部的结合过程，洛克说：“把一些由感觉和反思得来

的简单观念整合到一起，然后结合为复杂的观念”（《人类理智论》第 11 章第 6 节）。与之相反的过程是，“我们所有的复杂观念都可以解析为简单观念，即变为最初产生复合的原始观念”（《人类理智论》第 22 章第 4 节）。这就好比一台“心理机器”的工作，像砖块和砂浆筑墙一样，将一些简单观念结合为复杂观念。在结合操作过程中，运用的可能是联想原则的相似性和邻近性。但倘若如此，在这个操作过程中仅仅运用了众多原则中的两个，而一般而言这两个原则并不处于支配地位，“心灵……如果发现那样做很方便，就会任意地把它们结合成复杂观念；与此同时，那些实际上已经结合在一起的复杂观念也会变得松散，不再组合成一个单一观念，没有一个统一的名称。显而易见，心灵可以自由地对观念进行结合，某些时候那些互相已经结合在一起的观念，并不比那些以前被忽略的观念结合得更强”（《人类理智论》第 3 册第 5 章第 6 节）。可以看出，洛克认为复杂观念并不总源自观念的“处在一起”，而是可以“任意地”或者“自由地”形成——并不严格遵循联想原则。

观念连续性

在《人类理智论》第 4 版中，洛克专门新增加了“关于观念的联想”一章，创造了“联想”一词，后来该词甚至比他的理论更广为人知。但他对观念联想（他也使用观念“连结”）的关注似乎只限于病理学，即精神疾病上，也没有对联想原则进行明确命名和精确阐释。像他之前的霍布斯和他之后的休谟一样，也区分了两种不同的联想观念[2]——一种是观念“互相之间存在一种自然的契合或联合”，另一种则是“完全由于机遇和习惯得来，有些本来毫无联系的观念会在人的心灵中联想在一起，并且很难把它们分开。它们总是保持在一起，如果其中的一个出现在脑海中，另一个也会迅速呈现”。洛克对前者说得不多，但对后者作了详尽论述：(1)这种联结可以是主动的，也可以出于偶然（因此处于相同环境的人也会有不同的心理状态）；(2)深刻的第一印象，或者“时间强化（future indulgence）”（绝对强化？）“使观念联结得非常紧密，好像它们原本就是一个**观念**”；(3)一些不相容观念的联想“取决于心灵的原始构造，与生俱来”。洛克主要关注并实践修正**错误**联想——用于分析教育，而非心理学。随后，最杰出和最具影响力的联想主义者威廉·詹姆斯又带来了一次巨大转变。

威廉·詹姆斯：《心理学原理》

在心理学中，说到联想主义不得不提及威廉·詹姆斯的《心理学原理》

(*The Principles of Psychology*),该书第 14 章的标题就是“联想”。1878 年时詹姆斯计划两年内出版此书,但最终却花了他十二年的时间才完成[3]。威廉·詹姆斯(William James,1842—1910)可能是迄今为止,美国最杰出的心理学家。他比冯特小十岁,但当冯特 1875 年从苏黎世去莱比锡的时候,詹姆斯已经在哈佛得到 300 美元经费用于购置“心理学”仪器。“一般认为,冯特于 1879 年在莱比锡建立了世界上第一个心理学实验室,尽管冯特本人于 1875 年来到莱比锡不久就已经有了实验演示器具。简言之,冯特和詹姆斯 1875 年之后都设立了正式的实验演示实验室(并非研究实验室)”(Boring,1929:509)。19 岁时,詹姆斯在出国学习了一年艺术之后,进入哈佛劳伦斯研究院学习化学和比较生物学。两年后进入哈佛医学院。在 23 岁时,他同阿加西斯(Louis Agassiz)到亚马逊进行了一次自然考察。随后,又到德国学习了一年半医学。从此他涉足了多个领域的研究,当然也为认知科学作出了贡献。1872 年,他开始担任哈佛大学生理学讲师,1976 年成为生理学副教授。1880 年,他成为哲学副教授,并于 1885 年晋升为哲学教授。1889 担任心理学教授(正值其书完成)。

心理生活:思想

> 心理学是关于心理生活的现象及其产生条件的科学
>
> ——詹姆斯(《心理学原理》)

从认知科学的角度出发,詹姆斯关于心理生活或“思想”(詹姆斯说:“我用‘思想’一词是为了表达所有未加区分的有意识的形式。”)的概念具有如下关键特征: 24

> 1. 它是能够被意识觉察的。“从我们一出生就伴随着意识,它包含着大量的对象及各种联系。心理学首先就要假定思想本身是存在着的这样一个事实。心理学家也要首先关注我们时刻都在思想着什么的事实。”
>
> 2. 它是可内省的。“内省的观察是我们首先和主要的而且经常所要依赖的方法。”然而詹姆斯不似其后的一些人认为内省是难以控制的,他说:“内省是困难的、易误的……这种困难是一切观察都共存的困难。”
>
> 3. 它是私人的。“我的思想只同**我自己的**其他思想相关,你的思想也只同**你自己的**其他思想相关……我们所经历的每一种意识状态,都是一种特殊的意识、心灵或自我,出现在具体的特殊的我或者你之中。”
>
> 4. 它像“流动的河流”一样连续不断。内省过程及内容因观念按联想原则而彼此衔接,流动不断:“意识,不会在它断截的地点停留,不会

是一些割裂的片断，而是一种川流不息的状态……从此再提到它时，我们将称之为思想流、意识流或者主体生活流。”

5. 它总是关于什么（与目的相关）的。思想的构造成分是“观念”，而观念是关于什么的认识或者描述。

6. 它由进化得来。由于适应性，更高水平的认知功能才得以进化。

联想

背景

尽管詹姆斯像英国的经验主义一样，经常谈论心灵的复杂观念如何成为复杂整体的一部分，但他完全怀疑存在着复杂观念原则。詹姆斯同样有些矛盾地宣称，“参与联想的是对象，而非观念”。他说：“这样可以避免混乱，如果我们始贯如一地使用**联想**一词，只要这个词代表两个记忆事件之间的某种**作用**…在心灵中发生联想的并不是观念…或者只要这个词代表了**脑过程**间的某种**因果关系**”(Briefer Course:5)。但詹姆斯并没有清楚地说明：究竟产生哪种作用？也许我们可以用图 1.3 进行理解。

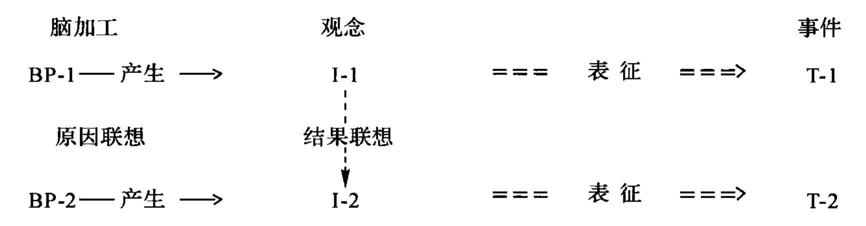


图 1.3 詹姆斯关于联想、观念和事件关系的观点

25 脑加工过程是联想动作的执行者。一种过程(BP-1)**引起**或者取得与另一种过程(BP-2)的联结。脑加工过程产生描述或者表征外界对象、事件(T)的观念(I)，因此不同事件得到联结——联想是脑加工过程的**功能**。观念是脑活动和事件的桥梁——由脑过程产生并同时表征外部事件。

无论詹姆斯怎样表达，他真正关心的是思维的时序性：即**心灵如何决定下一步应该想什么**？他认为，思维的顺序取决于联想原则，并列举了很多原则，包括相似律和邻近律。但詹姆斯不仅仅满足于对联想方式的**描述**上，而是常常强调如何进一步在神经水平上得到**解释**。例如，关于联想中的邻近律，他说：“不论我们怎样对其命名，他都是在表达一种心理习惯的现象，而**对其本质的解释，则认为他是由神经系统里的习惯法则产生；换言之，就是寻找其生理原因**”(《心理学原理》:561-2)。“在思维和经验中，相邻近的对象、事件产生联想，这种**心理联想律**其实是生理结构的功用，类如神经回路

可以最大化地在传导域得到传播”(《心理学原理》:563)。为了进一步解释,他还提出了两个重要并且具有前瞻性的神经原则,第一个只涉及两个脑加工过程,第二个则涉及多个脑加工过程:

(P1)

当两个脑加工过程被同时激活或继发激活时,一个过程如果再次得到激活,其会将兴奋传递到另一个过程。(Briefer Course: 5)

这个原则与我们将在第3章讨论的赫布提出的原则很相似。第二个原则是:

(P2)

26

脑皮层上任意点的激活程度,取决于其他所有发射点进入该点的刺激总和,所以激活程度的比例取决于:

1. 与该点相互联结的兴奋点的数目;
2. 兴奋的强度;
3. 与该点在功能上与之竞争的竞争点的数目,竞争点会争夺输入量。(Briefer Course:5)

第二个原则用一种想象的“神经元”给出图示,这对于我们的理解颇有助益,我们可将其称为“詹姆斯神经元”(见图1.4)。给出联想原则之后,詹姆斯把联想形式划分为**自发**联想和**随意**联想两种类型。

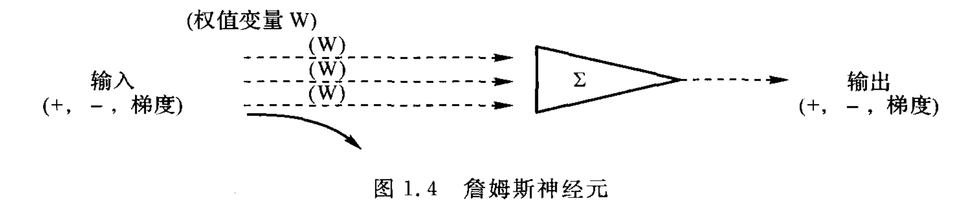


图1.4 詹姆斯神经元

自发联想

这里我们找到了这个现象的三种大致分类:(由相似性联想引起)整体回忆、局部回忆和聚焦回忆。

整体回忆

先前发生的事件与后来对之的回忆是非严格限制的联想。詹姆斯的一个宴会的例子能够对之作出很好的说明:

例如,如果a, b, c, d, e是宴会活动最后兴奋的神经束,称为活动A;l, m, n, o, p是在宴会后寒冷的夜晚走回家的兴奋神经束,称为活

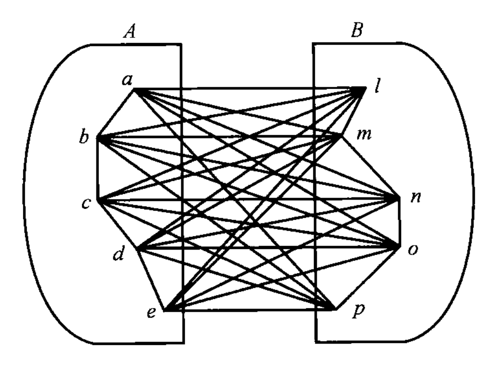
动B，那么想起A就必定也会唤起B。因为当a，b，c，d，e发生之后，会沿其路径输入到l，相似的又从l输入到m，n，o，p，并且这些后来的路径也会互相强化，因此在经验B时，“a，b，c，d，e”和“l，m，n，o，p”就会交互联结成为一个整体。图57的连线……表示传入B每个组成部分的输入，随后，当B作为结合了的整体而被意识到时，所有结合连接的强度也会增加。(Briefer Course：6)

自发联想	随意联想
1. 整体(部分)联想 [图57,58]	1. 回忆一件已遗忘的事 [图61]
2. 部分(混合)联想 习惯性 新近性 新奇性 情绪吻合 [图59]	2. 方法—目的推理 [图61]
3. 相似联想 [图60]	3. 所有问题解决的泛化

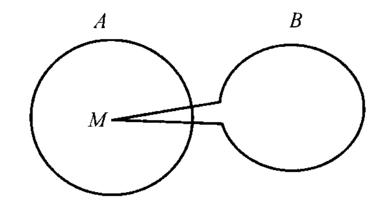
图1.5 詹姆斯关于思维连续性的分类

27 “我们刚刚描述的过程……如果没有遇到什么阻碍，将必然导致在回忆过去整个的、有着大量经验轨迹的内容时得到加强”(Briefer Course：6)(见图1.6(a)和1.6(b))。这样数量众多的和精细的联结并不是最常见的模式。

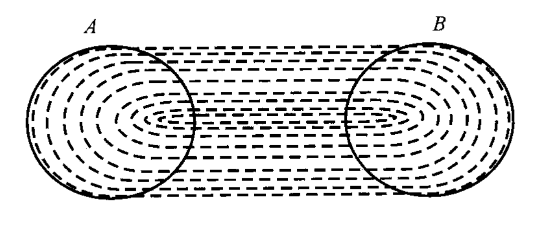
28



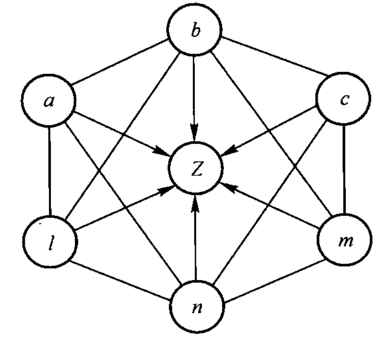
(a) 詹姆斯插图57：整体回忆



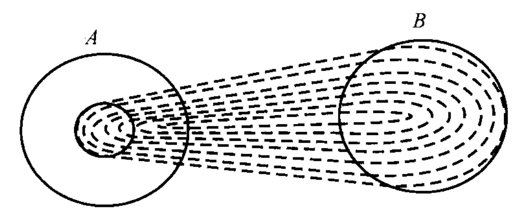
(d) 詹姆斯插图60：聚焦回忆



(b) 詹姆斯插图58：整体回忆



(e) 詹姆斯插图61：回忆与手段一目的推理



(c) 詹姆斯插图59：部分回忆

图 1.6　詹姆斯的思维连贯性示意图:联想的原则(来源:Briefer Course)

(a) 詹姆斯插图 57:整体回忆

(b) 詹姆斯插图 58:整体回忆

(c) 詹姆斯插图 59:部分回忆

(d) 詹姆斯插图 60:聚焦回忆

(e) 詹姆斯插图 61:回忆与手段一目的推理

局部回忆

局部回忆(见图1.6(c))是最常用的联想类别,只有部分过去经验是联想的结果,"我们回忆某一过去的事件,并不会回忆起其包含的所有项目,然后相应地判断下一个联想是什么。总是只有部分的组成项目,决定着其余联想的项目"(Briefer Course:7)。所以问题就变成,究竟哪些部分起决定作用以及为什么。詹姆斯的回答是,"起决定作用的是那些最能引起我们**益用**(INTEREST)的"(同上),"对于脑中发生的事项,其引起益用的规则是:**某一脑过程唤起别的脑过程,在其所有的联结中,这种唤起总是占据优势地位**"(同上)。詹姆斯总结了四种"益用"决定"思想复现"的原则。

(1)***习惯性*** 詹姆斯认为,习惯对于**经常出现的**过去经验里的元素有助于联想:"频繁性,在某种程度上是决定回忆的最强有力的因素。如果我突然地说出'swallow'这个词,读者如果是一位鸟类学者,将认为我说的是一种鸟,如果是生理或医学喉科疾病专家则会认为是指人的吞咽"(Briefer Course:8)。

(2)***新近性*** 詹姆斯给出了一个关于一本书的例子,这本书经常使他想到书中包含的内容,但是当他听到此书的作者自杀时,他再回忆这本书时,最先想到的是死亡。他总结:"意识倾向于唤醒他们最近以及他们最习惯(频繁)联想的东西"(Briefer Couse:8)。像往常一样,詹姆斯尝试在基础层次上解释这个现象:"特殊脑域的兴奋,或者脑中通常兴奋的特殊模式,给他们留下了一种趋势或者易再次触发的敏感性,但时间一长也会逐渐消失。而如果它一直持续,通过在别的时刻使它们发生反应,这些模式将易于唤醒它们的活动。所以**新近性**在经验中是决定意识回忆的最基本因素"(Briefer Course:8-9)。

29

(3)***新奇性*** 新奇性也决定着一种原始经验产生印象的强度或程度,"原始经验中的新奇性,与习惯性或者新近性产生回忆的作用相同"(Briefer Course:9)。例如,"如果'**牙齿**(tooth)'这个词出现在页面上,呈现在读者眼前,给他时间让他在无数种可能情况中,有五次机会回想任何情景,很可能他回忆出来的会是他在牙科门诊进行治疗的画面。他每天都碰触他的牙齿并咀嚼食物,每天早上清刷牙齿吃早餐以及剔牙,但是却不会或很少迅速地联想到这些,因为在牙科治疗的情景给他留下的印象非常强烈"(同上)。

(4)***情绪吻合*** 对此,詹姆斯写道:"第四个影响(在意识中)复制过去事件过程的因素是,我们现在的心情与回忆的内容**在情绪状态上吻合**。同样的一个对象并不能使人产生相同的联想,当我们正在忧郁的

时候并不能回想起愉快的记忆。事实上,当我们处于忧伤的心情时,我们也不会倾向于保持愉快的记忆……同样,那些性情总是乐观的人,当他们的情绪正高昂时,会发现他们不可能会有任何厌恶的征兆或者沮丧的想法表现"(Briefer Course:9)。

詹姆斯对这四个因素进行了总结:"**习惯性、新近性、新奇性和情绪吻合**,是一个人为什么在发散的意识中会因益用唤起这样的回忆而不是那样的全部原因。我们可以确定地说,**在绝大多数情况下,回忆起来的事情要么是出于习惯,新近发生的,要么是它生动鲜活,与当下情绪状态一致**"(Briefer Course:9)。

请注意,尽管詹姆斯标识了联想原则(associational principles, APs),并且为我们提供了例子,但并没有确切地进行**阐述**。这个联想原则究竟可能会是什么呢?詹姆斯并未专论,但如果这个原则是控制思想的时间过程,也许会是这样的:

(AP1)

如果主体正在对意识 A 进行注意,又因为 A 习惯(频繁多次)地与意识 B 联想在一起,那么主体接下来就会想到 B——除非这种联想被其他更强有力的联想原则覆盖。

(AP2)

在任何给定的时间中,最强烈的联想原则是最合适的、起作用的那个。

30

詹姆斯同样也没有把第四个(情绪吻合)与前三个原则区分开,然而它却可能是非常不同的,因为它似乎看起来并不能把任何特殊的意识(B)与另一个特殊的意识(A)联想到一起。它说明一组意识的整体比之部分,更容易回忆——这组意识在情绪维度上是相似的。

聚焦回忆或相似联想

这种类型的联想产生于事物间的局部相似,拥有同样的性质或关系(见图 1.6(d)):"我们假设益用注意选择的事物进一步精炼自己,并且突出强调是过去意识的某一部分,它变得太少了而不会再是具体事物的表象……只是具有抽象的性质或属性。让我们进一步假设,当其他部分已经消失,而这些却在意识中一直持续(或者在脑中,其脑过程依旧延续)。**这些剩下来的部分,当我们看到一些情景后它会单独进行联想**。这样,新的意识对象与消退了的意识对象便存在一种相似关系,这对意识我们称之为'相似联想'……。相似联想,其局部是同一的。当**相同的**属性出现在两个现象中

时，尽管只是它们的某一共同特征，这两个现象也会相似”(Briefer Course: 9-10)。这里，詹姆斯的例子是，说到月亮，然后想到的是煤气灯(颜色相似)，接着是足球(形状相似)。需要注意的是，月亮和足球它们自身并没有联结的性质关系。

· 随意思维轨迹

詹姆斯还试图从时间过程上对“无意识思维轨迹”进行扩展，用于解释“随意思维轨迹”:“到目前为止，我们已经假定从一个对象到另一个对象的过程是无意识的……就是自发的遐思和冥想。但在我们不断连续更迭的意识状态中，包含着与这些过程不同的、最主要的部分。它们明显地由动机或者注意兴趣引导，这部分观念的过程称作**随意活动**”(Briefer Course: 11)。像前面一样，詹姆斯同样从生理层次上进行了解释:“从生理上考虑，我们必须假设，一种动机的产生意味着有某些确定的脑过程正在持续激活……这种益用便是以我们假设的脑域的持续激活为基础”(同上)。在生理层次上，无意识和随意联想的最重要不同是，后者涉及神经过程的持续激活，而前者则没有。

31

随意联想一直是经典联想主义的一大阻碍，因为似乎理性、逻辑等过程总是时不时地优先于联想链。詹姆斯着手采取两个步骤处理这个问题。首先，他尝试用联想主义解释“回忆遗忘的事物”，然后尝试扩展这种解释到问题解决。詹姆斯使用问题以及问题解决解释随意联想:“在理论和实践生活中，存在一种更为敏锐的益用种类，对实现目的产生作用，使预期实现的目的具有确定的表象形式。一串连续的观念，在这种益用的影响下出现，构成通常所说的目的实现的**方法**意识。如果目的只是简单地呈现，没有及时提示出方法，那么再次搜寻实现目的的方法，会成为一个新的**问题**，并且找到这个方法也会变为一个新的目的……目的也就是指人强烈的欲望……但是有关它的本质……我们还没有任何明确的图景”(Briefer Course: 11)。所以，问题解决就被描画为方法一目的推理，詹姆斯随即对此作了延伸:“无论什么时候，发生了相同的事情，我们都会试图回忆一些已遗忘的东西”(同上)。“这种欲望牵扯和拉引出的方向，在我们的感受上总是正确的，但是其朝向的那个点却是不可见的。简而言之，某一**事项的缺失**，决定着我们的行为，与该事项出现时一样积极”(同上)。像前面一样，詹姆斯仍尝试在生理层次上重新描述这种现象:“如果我们尝试用脑活动的术语，解释潜存意识如何发挥作用，似乎要促使我们相信，与潜存意识对应的脑区域必然实存，只是极其微小且以无意识的方式实存”(同上)。詹姆斯认为这就是两类问题的共同结构，“回忆已遗忘的事情和搜寻给定目的的方法，它们之间的区

别在于前者已经成为我们经验的一部分，而后者则没有”(同上)。

回忆遗忘事项

就回忆一件已遗忘的事项来说，“这件已被遗忘的事项处于其他一些事件之中，但让我们感觉是一个缺口……我们因此会收集很多普遍与之相关的其他事件”。詹姆斯把回忆已遗忘事项的过程作了图解，如图 1.6(e)。

詹姆斯解释说：“回忆遗忘的 Z，我们最先感到的事实是它与 a，b 和 c 有关，之后我们回忆起其他的一些相关细节 l，m 和 n。每个圆圈中的字母亦代表着与 Z 的脑过程相连接的脑过程。激活 Z，一开始只有稍许张力，但是随着 a，b 和 c 的激活，慢慢地又浮现出 l，m，n 这些过程，它们以某种方式**全部**与 Z 有相关，它们的组合逐渐地倾向于激活 Z，用周围环绕的箭头表示，接着 Z 被完全激活回忆起来”(Briefer Course：12)。 32

问题解决：方法一目的推理

詹姆斯还思考了与回忆有关的问题解决。再回到图 1.6(e)(詹姆斯原初标识为图 61)，他说：“由于某种目的，首先在图中出现了 a，b，c，是唤起回忆线索的起始点。在这种情况下，随意注意不只是排除了一些不相干的线索，还紧紧抓住了那些他感觉更为相关且恰当的线索——用符号 l，m，n 表示。所有这些，最终积累充分，共同指向了 Z。在意识领域，Z 的最终兴奋过程等价于问题解决。这种情形与前面讲的(回忆遗忘事项)唯一区别在于，Z 不必要有自身的潜兴奋，a，c，l，n 等的共同协作才是首要的”(Briefer Course：12)。接着詹姆斯把他的结论概括(有些轻率)为：“**从猜测报刊上一些难以理解的事物，到标绘出帝国的政策方针，都是这样一种过程**。我们相信，脑的自然法则会自发地给出我们恰当的想法”(同上)。

詹姆斯从来没有怀疑，人们所使用的推理是否都是他所描述的方法一目的推理(也没有怀疑是否他所描述的方法一目的推理都是正确的)。思考权量支票簿的问题：一个数加上一组数字列，再减去另一些数，然后比较结果等等，这些过程都是(詹姆斯的)方法一目的推理吗？当加上一列数，并进位一个“1”时，我们也是随意处理这种联想并等待必要的联想突然进入意识？似乎不是。此外，在这种情况下，问题解决最好用目的来描述，而不是方法——方法是作为计算原则，这里我们又看到了霍布斯早先对工作时联想过程与“计算”的区分。也许詹姆斯的理论只适合于前者，他推广到所有的问题解决就是错误的了。值得注意的是，詹姆斯怀疑这种推理能够得到完全解释的可能性：“分析各种心理诉求过程的细节并不是我的目的。在科学研究中，我们可能会找到大量的例子……我们的审查机制中并不存在什么规则，都是直接通达目标结果的……最后得出的发现结果仅仅是与预期

相一致，并非是因规则（联想）作用。最终需要脑区自身主动地直接指向正确的途径，否则我们将始终是在黑暗中探索……我们总获益于脑内有关相似性的无意识加工”（Briefer Course：12-13）。

33 詹姆斯同时怀疑完全的脑科学是否可能：“唤醒记忆的基本过程除习惯规则外，别无其他。生理学家距离实现从一个假设已激活的细胞集追溯扩散到的另一细胞集，还很遥远，或许永远也达不到”（Briefer Course：13）。詹姆斯又一次返回到了神经层次，总结如下：“总结起来，**三种联想方式两两间的差异，可将自身简单的还原到脑区特定位置上神经带激活数量的不同，那些神经带支持对即将出现在意识中的内容进行回忆**”（同上）。“**心理材料呈现**的顺序仅仅出于脑生理原因”（同上）。可见，詹姆斯所有分析的要点是用联想原则**描述**心灵中呈现的思维序列，然后尝试用神经生理学原理**解释**思维序列如何呈现。

1.5 经典联想主义的终结

从 19 世纪末到 20 世纪初，至少有三个方面的因素使经典联想主义在心理学领域开始衰落。每种因素都与偏离了英国经验主义的内省观念相关。首先，**内省方法**自身的破产。在 19 世纪末期就因为没有方法解决这个问题而出现众多争论。自冯特在莱比锡（1879）建立实验室以后，心理学成为一门科学，感知等心理活动就很少或者不需要写意的文字风格对其描述。认为那是繁杂堆砌且无法证实的轶事，所以就到了结束“紧张内视”的时候了。第二，有很多相关因素随后转向了**行为主义**和**刺激－反应**心理学。也许最初的发展是由于艾宾浩斯（Ebbinghaus）对无意义音节学习所做的有关刺激－反应的研究工作。这可看作是联想主义原理在实验室中的第一个应用，同时也是使学习、记忆和思维在作为科学的实验心理学中迈出的第一步。接着是桑代克（Thorndike，1911）对动物学习的研究，为华生（Watson，1913）和从 20 世纪 30 年代到 40 年代的行为主义运动开辟了道路，同时行为主义也融入了巴甫洛夫（Pavlov，1927）有关条件反射的重要研究（下一章阐述）。第三，哈特莱和詹姆斯致力于呼吁要在**神经**层次上解释心理现象。这
34 种解释策略，伴随神经科学（与“生理心理学”）以及神经系统技术和理论（高尔基、卡哈尔、谢灵顿）的发展，希望集中于用（纯粹）心理学原则所能研究的问题（第 3 章阐述）。随着内省方法的终止，关于对象的内省－观念的研究也便终结了。有关心智的新要素变为刺激和反应，它们的神经基础——完全

与内省无关。随着观念被刺激和反应取代,内省也被实验室实验取代,同时还增加了很多重要的概念,如强化、奖励和条件——这些在英国经验主义那里都是很少提到的。

注释

[1] 正如我们看到的,詹姆斯的明确立场是,被联想的是事物(记忆中的)而不是观念。

[2] 对于洛克而言,与其后的休谟不同,"观念"中包含所有的心理内容:感知观念和反射观念。

[3] 引自威廉·詹姆斯《心理学》第 16 章"Briefer Course",标示的数字为"Briefer Course"中的页码。

【思考题】

什么是联想主义?

"联想主义"原则的一般表述是什么?

"完全联想主义"与"混合联想主义"的含义分别是什么?

感知受联想支配吗——为什么/为什么不?

大多数联想主义者认同的联想的两个原则是什么?

联想主义者的原则应当解释的两种主要心理过程是什么?

什么是联想的邻近律?

什么是联想的相似律?

什么是联想的相对律?

什么是联想的因果律?

洛克与詹姆斯

洛克对联想主义的主要贡献是什么?

根据洛克的观点,心理学应当研究什么?

根据詹姆斯的观点,心理生活的六个主要特征是什么?

35

詹姆斯的主要关注于联想的什么过程?

对于詹姆斯而言,基本的解释层次是什么?

詹姆斯对于思维可能的完整解释层次的观点是什么?

涉及两种脑过程共同激活的基本控制联想原则是什么?

涉及多种脑过程共同激活的基本控制联想原则是什么?

"作为事物间效果关系的联想"是指什么?

"作为脑过程间原因关系的联想"是指什么?

益用在无意识联想中的作用是什么?

詹姆斯对于局部回忆分类的四个原则是什么?
给出每一原则的示例(詹姆斯所做的)。
试着将其作为联想原则表达清楚。
相似性联想是什么? 给出詹姆斯的例子。
詹姆斯说相似是指局部相同,是什么意思?
随意思维与无意识联想的区别是什么?
詹姆斯将随意思维划分为哪几类?
回忆已遗忘的事项与方法一目的推理的区别是什么?
哪种推理对詹姆斯的观点提出了挑战,为什么?
哪三个方面的因素导致了经典联想主义的终结?

【推荐读物】

概述:Warner(1921)对联想主义作了完整的描述,虽然出现的时间较早,但奇怪的是,该书并没有提及詹姆斯。Boring(1929)第 10 章阐述了英国经验主义,第 12 章介绍了密尔父子和培因的观点。Marx and Hillix(1963)第 6 章,介绍了联结主义和早期的行为主义,将行为看作是刺激和反射间的联想。Anderson and Bower(1974)是一部导论性著作,与当代人们的观点更为接近。

36 关于洛克的更多研究见 Cummins(1989)第 4 章,关于洛克观点的更多论述可在 McCulloch(1995)第 2 章中找到。更多有关詹姆斯的讨论见 Flanagan(1991),其中第 2 章有关于詹姆斯心灵哲学以及从认知科学视角审视心理学的精彩讨论。对于其他经验主义者的阐述,如休谟,可见 Wilson(1992)。Young(1970)第 3 章,以当代视角对培因的观点作了一些探讨。Beakley and Ludlow(1992)第 IV 部分,为联想主义著作精选。Hunt(1993)第 3 章简明地评述了经验主义和唯理主义心理原则,第 6 章则对詹姆斯作了一般性的讨论。

2 行为主义与认知主义

2.1 引 言

我们在前一章中简要回顾了联想主义的兴衰。在联想主义与心智计算 37
理论(计算或联结主义)之间是行为主义或者说是刺激一反应理论(S-R)的兴盛时期,主要在美国、英国和澳大利亚广为传播(可以看作是在英语国家兴起的运动)。当然,在此期间也出现了许多其他重要理论,如弗洛伊德的无意识学说,格式塔学派的内部知觉组织理论,动物的社会行为研究,以及皮亚杰对于儿童认知发展阶段所做的工作等(可以看作是发生在欧洲,主要由德语国家发起的运动)。在这一章中,我们主要讨论行为主义缘何兴起,它包含的一些基本理论假设,以及为什么会衰落等问题。从中可以发现,行为主义存在的一些不足,以及它为认知主义或者信息加工心理学的诞生和心智计算理论的最终出现所作出的贡献。

2.2 行为主义与刺激一反应心理学的兴起

探索人类心智的新时代

——《纽约时报》(1942)

大西洋彼岸的美国,从威廉·詹姆斯直到二战结束,在探索人类复杂认知过程方面一直存有空白。尽管这并不是绝对的,但是却不得不

说美国的认知心理学一直被行为主义、无意义音节和老鼠所左右。

38 ——纽厄尔和西蒙(Newell and Simon,1972:874)

1910年到1913年期间是行为主义形成的关键时期。到1911年左右,几乎所有心理学家都回避谈论意识及把内省的心理**内容**作为研究对象,而是转向研究心理活动的外部行为。因此,内省的方法自然地也不再是主要的心理学研究手段。1911年,美国心理学会专门召开了一次论坛,讨论从研究意识到运用行为进行解释、描述和控制的这种研究转变。这场运动的经历者安吉尔(Angell)写道:"毫无疑问,一种新的运动正在展开,它的兴趣集中于讨论意识活动的**结果**,而不是**过程**本身。这在动物心理学上其实一向如此;但是在研究人类的心理上似乎还应用不多。用一个较为恰当的词汇说明它的兴趣对象,那就是'行为';分析意识从根本上要研究人的行为,而不是反过来"(Angell,1911:47)。

巴甫洛夫

巴甫洛夫(I. P. Pavlov,1849—1936)继承和发展了谢切诺夫(I. Sechenov)的思想(尽管并不是他的学生)。谢切诺夫是"俄国反射理论的开创者",提出了著名的**脑反射理论**(1863),尝试用纯粹的生理活动揭示心理现象,"所有的心理活动毫无例外,除了涉及复杂的情绪元素(后面阐述)都可通过反射而得到。因此,所有的意识活动(通常称之为随意行为),如同所有的行为动作一样,都是一种反射。'反射'是一个最能精确概括的词汇"(1863:317)。巴甫洛夫的工作可以说是整合了所有传统反射理论,我们将在后面看到笛卡尔和联想主义都是其理论学说的先驱。巴甫洛夫于1904年因他在消化领域的贡献获得诺贝尔奖,然而他最著名、影响最广、贡献最大的发现——经典条件反射,却与诺贝尔奖无缘。

"经典条件反射"包含几个部分:狗吃食物(非条件刺激:US)会引起唾液分泌(无条件反射:UR)。在给予狗食物(US)之前给予铃声刺激(条件刺激:CS)。重复多次条件刺激以后,单独的条件刺激也会产生无条件反射,形成条件反射(CR)。

39

食物(US)→分泌唾液(UR)
铃声(CS)+食物→分泌唾液(UR)
铃声(CS)→分泌唾液(CR)

巴甫洛夫还发现了一些反射原则:条件反射出现在无条件反射的前面,

而不是之后。条件反射同样也可以由与原来条件刺激相似的刺激引发，称为“泛化”；有机体通过辨别学习，有选择地对某些刺激作出反应，而不对其他刺激作出反应，称为“辨别”；当条件刺激多次重复而不伴随无条件刺激，条件反射将逐渐削弱直至消失，称为“消退”；然而假如人们让动物单独待一会儿，那么条件刺激又会引起条件反射，称为“自然恢复”。巴甫洛夫认为，所有的行为都可由先天具有和后天习得的反射进行分析，并尝试寻找构成反射的所有生理机制。如他所说：“一方面，非常有必要找出所有的先天反射并对其进行系统化，它们是最基本的，永恒不变的。在此基础上，有机体才可以获得大量的后天习得反射……另一方面，同先天反射一样，需要找出习得反射所遵循的原则和机制……只有掌握了所有的反射活动，我们才能逐渐揭示高等动物生活的全部秘密，这才是科学的分析方法……我坚持认为，首先描述和找出所有的最基本的先天反射，对于逐渐理解动物的所有行为是必要的”(Pavlov，1928：281-3)。巴甫洛夫似乎认为这种解释可以扩展到动物(包括人类)的所有行为，“整个心理机制就是由精细的基本联结[如前所述]和一连串的联结链所构成的”。

华　生

1913年，《心理学评论》刊登了华生(J. B. Watson，1878—1958)的《行为主义者心目中的心理学》一文，被看作是行为主义兴起的宣言。1943年，一些杰出的心理学家称这是《心理学评论》已发表的最重要的文章。华生早期从事动物行为研究，自然地很少与意识和内省打交道。这倒很符合华生厌恶心理学偏重意识和内省的时代(如，詹姆斯和冯特对心理学研究的态度)，同样对心灵与物质的关系(心一身问题)这一问题不屑一顾。他说：“那些自古以来的哲学空论的残余，并不为行为主义者增加一丝一毫的困惑，就好像它们并不给物理学家增加困难一样。思考心一身关系问题既不能在客观上得到观察，也不能找到方法予以证明，找到任何解决这一问题的方式。”在动物实验室中，华生进行了一项提高鼠类学习能力的实验，成为一项重要的研究范例。之后，他尝试把这种方法扩展到研究人的心理现象上。在《行为主义者心目中的心理学》开篇，华生就直接指出：“在行为主义者看来，心理学是属于自然科学的实验科学分支。它的理论目的在于预测和控制行为。内省并不是其方法的主要部分，它的科学价值也不依赖于这些资料是否容易运用意识的术语来解释。行为主义者致力于获得有机体所有反应活动的统一图式，认为在人类与动物之间并没有绝对的区分。人类的行为，不管多么

40

精细、复杂，也不过是行为主义整个研究图式的一部分”(1913:158)。从这段简短的引文中，可以这样概括行为主义的主要观点：

1. 心理学是一门客观的自然科学，运用的也是客观的自然科学方法；

2. 心理学的目的是预测和控制行为；

3. 人类与动物的心理并无本质区别；

4. 心理学研究应排除内省和意识。

与哈特莱和詹姆斯一样，华生也寻求生理学的解释。从行为主义者的角度看，“心理学的所有发现，都取决于与生理相关的功能结构，并最终都可以用物理—化学的术语得到阐述”(1913:177)。最初，华生用“习惯”解释他的行为理论，但在1916年时又吸收了巴甫洛夫提出的条件反射理论，作为理解习惯单元的正确方式。他关于习惯的一些观点会让我们感觉有些不可思议(但还不是最糟糕的)，例如，他声称“思维”(并不清楚行为主义者为什么还使用这个词汇)与大脑无关，只是由“喉部感觉—运动过程”所构成。

桑代克

桑代克(E. Thorndike，1874—1949)最初的研究兴趣集中在儿童学习和教育，但是因为缺乏实验被试而转向研究动物学习，实验室就设在詹姆斯的地下室中(哈佛大学并不为他的儿童研究提供实验场所)。他早期对于操作性学习所做的工作，都收录在《动物智慧》(1911)一书中。基本的实验装置是一个“迷箱”，动物(主要是猫)如果做出恰当的反应会得到奖励(食物)。41 例如，实验开始时，猫会在迷箱中做出大量试图逃离迷箱的无效反应，但只要它偶然踩到了踏板就能逃出迷箱。经过如此多次后，猫就学会了这个反应，它会直接做出反应，打开迷箱，吃到外面的食物。但如果取消给予奖励，那么它的这种反应将会逐渐消失。学习就是反复尝试与试错、奖励与惩罚的过程。桑代克描述他的实验目的，是教会动物“使用它们的大脑”，但是他的方法却是(不太仁慈)“让一只饥饿的猫，偶然地学到一个反应，逃出箱子得到食物”(Hilgard，1987:190)。早期对桑代克的批评，如柯勒(Kohler)所说，应该让动物处在自然条件下才可进行研究，而不应在实验室中，情境条件的变化会影响动物做出的反应。桑代克提出了两条反应律，认为可以用于解释所有动物(包括人类)的行为，也就是他所谓的“效果律”和“练习律”：

效果律 在特定情境中所做出的行为，如果获得了满意的效果，那

么这个行为就会与情境产生联结，当情境再次发生时就容易引起该行为。相反地，如果产生痛苦的效果，以后再出现同样的情境时则不会做出该反应（积极强化和消极强化）。

练习律 刺激与反应之间联结的强度、效力和持续性，在其他条件不变的情况下，与在那个情境中使用的次数成正比。

不难看出，第二条反应律受到了詹姆斯的影响。桑代克在他随后的《人类的学习》(1929)中扩展了这些反应律，用以解释人类的学习行为，认为人类的学习也是刺激(S)与反应(R)之间的联结。在每一个S-R链中，S都具有产生R的概率；学习是概率的增加，而遗忘则是概率的减少。桑代克还发展了一种关于学习的神经理论，认为学习是突触（谢灵顿于1906年发现）间建立了新的联结。因此，桑代克也常称自己是“联结主义者”。

斯金纳

华生之后最著名、最有影响的行为主义者，毫无疑问是斯金纳(B. F. Skinner，1904—1990)了。他也不承认任何心理状态和过程的存在，甚至走得更远。受逻辑实证主义的影响，他拒绝任何形式的“假设”实体。他早期(1931)甚至不接受“反射”的概念，认为它也是一种假设的实体，更愿意把它看作是“对刺激和反应稳定关系的一种简单描述”。斯金纳认为，心理学的目的是运用实验的方法找出在特定环境中引起行为的原因。用他常用的术语来说，就是把有机体看作是“变量集合”——包括环境、“自变量”和“因变量”。心理学的主要研究目的是找出这些独立于生理现象的变量间的作用关系——尽管与之前詹姆斯的观点一致，认为心理问题最终都要归结为具有那些功能关系的生理机制。 42

斯金纳在其第一本重要著作《有机体的行为》(1938)[1]中描述了这样一种心理学研究纲领：心理学实验要分析动物和人类操作行为中，组织化了的“强化列联(contigencies of reinforcement)”。**强化列联**包含：行为发生的背景、强化反应和强化物三个部分。斯金纳在实验中使用了一种（被戏称为）“斯金纳箱”的装置，其方法是：(ⅰ)把饥饿的白鼠放在箱中，让它自由活动；(ⅱ)控制环境中的各种变量；(ⅲ)如果老鼠一个无意的、简单而明显的动作（例如，按压杠杆或拧转钥匙），就可以得到食物；(ⅳ)记录随后白鼠重复此动作的次数和频率等几个步骤。这样的设计选择主要是为了简化对强化列联的分析。斯金纳还区分了两种不同的学习行为：(1)反射（或者“应答”）行为，如巴甫洛夫实验中的唾液反射；(2)操作行为，如白鼠主动学会按压杠杆

或拧钥匙。个体的操作行为与物种的自然选择其实是一回事。有机体无意的一个新动作,如果获得积极的效果,得到奖赏,那么这个动作重复出现的可能性就会增加(正如"基因库"中出现了一种新变异,如果能获得较多的成功,那么它被遗传保留并在人群中扩散分布的概率就会增加)。斯金纳的目的就是希望能够在强化影响条件下,最终对操作行为的控制原则系统化。在他看来,所有的行为无非都是个体经历的所有强化过程与基因共同作用的结果。

1957年,斯金纳出版论著(《语言行为》)尝试证实他的观点。从笛卡尔开始,语言就被认为是人类特有的能力,所以语言也就必然成为检验行为原则适用范围的合适选择。然而,斯金纳的观点并没有得到新的实验证实,也没有涉及语言本身。仅仅尝试建立一种解释语言行为似乎合理的适用框架——准确地说,是如何在特殊环境条件下获得语言的动作。他的这种尝试并不成功,随后我们将看到乔姆斯基对他的质疑。斯金纳的荣誉仍然集中在动物领域。

2.3 对行为主义和刺激—反应心理学的挑战

43

20世纪50年代,**行为主义**在美国心理学(及相关的认知科学)领域中的统领地位宣告结束,开始逐渐向**认知**、**信息加工**范式转变。与此同时,出现了四种重要的趋势:首先,逻辑实证主义由于它的科学方法理论架构的非现实性和局限性而几乎被哲学界抛弃,其他很多学科也受到了影响。其次,行为主义的研究严重脱离了现实生活,大量论文只有该领域内的人才能看得懂,正如当时一位心理学家写道:"因为这种趋势,使得人们觉得心理学的研究对现实生活的意义不大,随之人们对心理学的兴趣也急剧降低"。第三,在此期间,行为主义的基本假设也受到了严重批评,批评的重点是指出,它没有说明行为的内在结构(无法用行为主义和刺激—反应心理学解释),并忽略了有机体对行为的影响。最后,在研究课题中出现了较多使用计算和信息术语的趋向。现在我们着重讨论第三点,最后一点将留在下一章讨论。

拉什利

1951年,拉什利(K. Lashley)发表文章批评行为主义的一些基本原则,产生了重要影响。他讨论的焦点是行为主义所不能解释的具有序列顺序的

行为："具有序列顺序的行为是一个重要问题；存在着一种具有时间等效性的产生行为的图式，它决定着具体行为的发生顺序，或作用于其自身或作用于与其相联结的其他行为"(1951:122)。

在行为主义者看来，按照刺激一反应理论，一种行为就是一串刺激一反应联结链，每一反应由上一刺激引发并引起下一个反应，如凭记忆用乐器演奏一串正确音符：

S1→R1/S2[第一个音符]→R2/S3[第二个音符]→…

拉什利批评了这种观点，他认为如组织语言、协调动作、整合时间与节律等行为都是 S-R 联结链不能解释的，"从节律活动或空间定向行为中，我认为可以推出在神经组织中必然存在着一种所有神经元相互联结、精细而复杂的系统，把大量不同类型的空间效应器得到的结果整合为统一整体"(1951:127)。作为替代概念，拉什利(如詹姆斯一样)转向了神经系统，认为神经系统对于输入的感觉刺激材料有着自己的组织原则，起着主动的作用，而不像旧 S-R 理论那么被动，"按照我的理论原则，输入刺激并不是一个静态的、不变的系统，而是总处于兴奋和被组织的激活状态。在一个健全的有机体中，行为是输入的特定刺激经内部兴奋条件的作用而产生的结果"(1951:112)。 44

乔姆斯基

对斯金纳最有力的批评并不是来自心理学家，也不是生理学家，而是一位年轻的语言学家乔姆斯基(Noam Chomsky)。他因为对斯金纳 1951 年出版的《语言行为》提出了质疑而受到广泛关注，此后行为主义也便不再被认为是心理学的研究框架。乔姆斯基还开创了一种新的认知取向，"乔姆斯基的这篇评论论文，也许是自华生行为主义宣言后最有影响的心理学文献"(Leahey,1992:418)。乔姆斯基对斯金纳(最切中要害)的批评主要是针对他的语言分析理论，当然他的批评并不限于此。乔姆斯基还质疑了斯金纳把人类活动与动物行为可用同一框架解释的充分性，尤其是比较生态学强调的那部分。在结尾部分，乔姆斯基从生成语法角度提出了一种新的框架。

对斯金纳普遍框架的批评

乔姆斯基首先列举了斯金纳提出的刺激、反应("应答"、"操作")和强化等概念，然后指出，尽管这些概念在实验室背景中能得到明确定义，但却并不能用来解释人在现实生活中的行为，"认为如此严密的科学理论完全可以

广泛应用，这不过是一种幻觉。实际上，即使描述人在现实生活中的语言与描述实验室中行为的语言相同，其意义也是不一样的，至多仅有一些模糊的相似性”(1959:30)。然后，乔姆斯基用了三段内容支持他的批评，完全切中要害。

对斯金纳语言行为的批评

45　乔姆斯基首先列举了一些斯金纳用于描述语言行为的概念，如“自发的”、“祈使的”、“娴熟的”言语等，然后说道：“这些概念几乎涵盖了斯金纳体系的主要内容，我之所以逐一讨论这些概念是为了说明：如果我们认为它们只具有文学价值，那么斯金纳的描述则完全没有涉及语言行为的任何方面；如果认为它们是隐喻式的，那么比之传统的观点也没有任何改进”(1959:54)。

在审视了斯金纳刺激、反应和强化的概念之后，乔姆斯基说，斯金纳致力于在有限的、严格控制的实验条件下发现一些动物的行为。即便如此，也只能观察到那些在可观察的条件下引发的动物行为，而遗漏了动物内在的理解行为：“描述复杂有机体的行为，自然会想到……除了需要了解外部刺激之外，还要知道有机体内部结构的知识，以及有机体如何处理信息，怎样组织其行为等内容”(1959:27)。“仔细研究斯金纳的书……可以发现……强化理论者在实验中取得的成果，尽管非常令人信服，但是他因此将结论推广到复杂的人类行为，未免过于草率和肤浅……之所以不能把刺激一反应理论用来作为一种测量手段来解释语言行为，其主要原因是，我们对语言行为这种非常复杂的现象所知甚少”(1959:28)。

结　论

拉什利和乔姆斯基对行为主义批评的意义在于，尽管行为的各个方面都能够用环境、刺激和反应以及内部刺激一反应链进行控制和解释，但是，**解释复杂生物体的复杂行为，却需要考虑到有机体、它的内部组织及其操作原则的作用**。从这个角度来看，信息理论和数字计算机显然是完全符合要求的。

2.4　认知主义：信息加工心理学

行为主义的衰落引起了多种竞争范式的出现，但认知主义迅速成长起

来，它主要源自信息理论和(将要成为独立学科的)计算机科学而非主流心理学，“在很长一段时间的竞争中，一种最重要的认知取向从数学和电子工程领域出现了，尽管它们与心理学及其相关问题联系较少，甚至没有任何关系”(Leahey,1992:397)。数字计算理论是本书第二部分的主要内容，现在我们主要关注的是人们从20世纪50年代到60年代所做的一些初步且重要的工作。也正是在这一时期，人们开始使用并界定了信息加工的概念。 46

信　息

“信息”具有多种含义，最早通用的用法很难进行定义(因此只是狭义的技术概念)，有点类似于：关于一件模糊事件的，或者什么东西是什么的准论点，这样的例子可以在《信息年鉴》中找到。根据这种通用用法，所得到的信息既不为真，也不为假，并且没有明确的测量方法。有时这个概念又称作“语义信息”，因为它是“关涉”于某事物的(语义)。与信息的这种通用用法相对的是技术上的使用，由申农在20世纪40年代早期提出，包含如下含义：信息就是减少可能性、无知状态或者不确定性。这种定义关注的主要是信息的测量，测量的标准单位是bit(二进制数)。1 bit的信息就是减少关于某一情形的不确定性或无知程度的一半，如抛掷一枚硬币，就是减少有两种变化状态的可能性(正面和反面)到仅有一种。2 bit信息是减少涉及四种可能情形到一种(第一个bit：从4到2，第二个bit：从2到1)。一则**消息**(信息)从**发送者**出发，经由**通道**传递到**接收者**，通道具有一定的容量和某种程度的噪音。

纯粹的信息理论主要关注信息的测量，以及在各种不同容量和噪音程度的通道中信息的传送。然而，对于早期的信息理论心理学来说，信息加工过程却是最主要的。更重要的是，早期信息心理学如Neisser(1967)，倾向于关注感觉的输入(视觉和听觉)，在第11章中只用了一章的篇幅讨论“高级心理过程”，如记忆和思维等。部分原因在于，信息的技术概念能够对感觉提供相当直接的应用：环境“发送”一条“消息”，消息在“噪音”背景下通过感觉“通道”，最终被有机体接收并理解。有关这种观点最早的流程图表达见图2.1。然而，当转向记忆和思维时，如何应用信息的技术概念就完全不能清楚说明了，所以信息的概念逐渐地在不知不觉中最终被“信息加工”所取代，成为一个双关语。 47

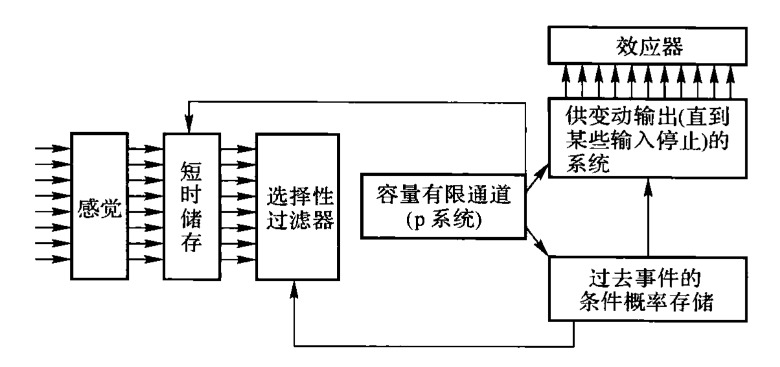


图 2.1 最早的信息加工图解之一(来源:Broadbent, 1958:299,图 7)

米勒:“神奇的数字:7±2”

乔治·米勒(George Miller)热衷于早期信息理论及其在心理学中的应用。他的早期著作《语言与交流》(1951)及众多论文,尤其是《神奇的数字 7……》(1956),“激发了人们对注意和记忆容量限制的研究兴趣,并掀起研究信息一加工心理学的第一波浪潮”(Leahey,1992:406)。如米勒自己所言:“信息的概念已经证明在研究人的识别和语言能力中具有重要价值;预示了研究学习和记忆的广泛前景”(1951:42)。米勒在他那篇最有影响的论文中,开篇就写道:“我的问题是一直因一个整数而困惑。这个数字已经缠绕了我整整七年,就是它使我得出的实验数据都像是我故意弄出来的,使我的文章遭受了众多公共期刊的退稿。”在论文的正文部分,他详述了十二个实验,并得出如下两点结论:“首先,绝对判断的长度和短时记忆的长度在我们接收、加工和记忆信息量上有严格的限制”(1967:41);“第二,在人类心理学中,再编码的过程是非常重要的,应当受到更多的关注”(1967:42)。下面我们介绍一些关于这两点结论的细节。

线性刺激的绝对判断

48 米勒检验了关于四种绝对判断的学习。例如,当主体被要求辨认出分派给他们的数字的读音时,得出这样的研究结果:“当给予两个或三个数字时,被试从不会弄混淆。当有四个不同的数字时,错误开始出现,但也比较少。但是,当有五个或者更多的不同数字时,混淆错误就频繁地发生了”(1956:18)。在关于听觉、味觉和空间位置上也有相似的判断结果报告。米勒得出:“或因为学习,或因为先天具有的缘故,在我们的神经系统中似乎存

在一些固有的限制。这些限制制约了一般变化幅度的通道容量。在现有证据的基础上,似乎可以明确地说,对于线性判断,它的容量有限且很小,并不因一种简单感知到另一种简单感知而发生变化"(1956:25)。他指出,这个容量就是7加减2。

再编码

既然我们的记忆长度限制在这样一个小数目的信息"组块"里,那么能够在每一"组块"中填充更多的信息就非常重要了,这个过程称作"再编码"。米勒列举了扩展对二进制数进行记忆的组块研究,他本人能够以5:1的编码方式,没有错误地记忆40个二进制数,"再编码是一件我们能够掌握的,对增加记忆的信息量颇为有效的方法"(1956:40)。

在这一领域,米勒的论文具有重要影响,正好与乔姆斯基对结构语言的批评(同时也是对行为主义的批评)出现在同一时期。它不仅概括了各种实验得出的结果,而且使人们认识到了有机体内存在一种内在的信息一加工结构——这也正是乔姆斯基在语言学领域所提议的。

米勒、加兰特尔和普利布拉姆的《行为的规划和结构》

前面提到,乔姆斯基(1959)批评斯金纳以及绝大多数行为主义者忽视了"有机体处理输入信息和组织其行为的内在知识结构"。行为主义和刺激一反应理论认为,内在的S-R链条"调解"输入刺激与输出反应,但我们同时也看到,拉什利(1951)和其他一些人提出最值得关注的是行为的内在结构 49 而非S-R链条。尤其是,当有机体有一个目标,并需要知识、计划、策略和手段来实现这个目标,这就部分地需要具备一种层次组织结构,而不是线性的链条。米勒、加兰特尔和普利布拉姆在《行为的规划和结构》(Miller, Galanter, and Pribram, 1960)中就表达了这种观点,对复杂行为提出了重要分析。

疑难

米勒、加兰特尔和普利布拉姆(后文简称MGP)认为,一个人在通常的一天里,总是围绕着两个中心概念活动:期望或者想象将要发生什么,以及计划怎样应对。他们说,这些概念在行为主义框架内进行分析几乎是不可能的,事实上,"近年来,很多人已经开始怀疑反射理论太过简单,它的要素也太过简单"(1960:22-3)。与"反射理论"相对的是"认知状态","他们(认知主

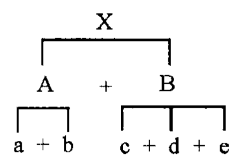
义者)认为一个事件对行为产生的影响,取决于这一事件如何在有机体内表征自身以及与这一事件相关联的所有其他事件的总体。他们确信,刺激与反应之间的任何相互关系,必然受到关于有机体所处环境的,且环境被表征组织化了的概念和关系系统所调解。人类——或者其他动物——具有一种内在的表征,它是关于环境整体的模型、图式、模拟、认知地图或镜像(image)"(1960:7)。然而,尽管在知识与行为之间的空缺小于刺激与反应之间的空缺,但空缺依然存在,必须得以填补,"很多心理学家,包括作者本人,都对这种在认知与行为之间的理论真空而困扰。这本书主要就是讨论……如何填补这个真空……描述有机体内在的整体表征如何控制行为"(1960:11-12)。

解决途径

MGP 提出的解决方案主要包含五个部分:区分关于克分子和分子的分析层次、镜像、规划、执行和 TOTE 单元。

行为的层次结构

解决这个疑难首先需要了解"行为的组织层次"。解释和理解人类大部分的行为,能够并且必须同时在不同的层次组织上进行分析:X 由 A+B 构成,A 由 a+b 构成,B 由 c+d+e 构成:



反映在语言中的一个例子:X 可能是一个句子,A、B 是短语,a—e 是单词,还可以继续认识到单词由词素构成,而词素是由音素构成。非语言行为同样也可以这么看待,X 也许是指发动汽车,A 是指踩油门,B 是拧转钥匙,a 是把腿放到油门踏板上,b 踩下,c 是把钥匙从口袋里掏出,d 是把钥匙放进启动器中,e 是拧动。至少是从托尔曼(Tolman)的工作以后,心理学家包括 MGP,称较大的、较高层级的行为单元为"克分子"单元,而称较小的、较低层级的行为单元为"分子"单元。行为一经被认为是具有潜在复杂性和多层次性的组织,我们面对的问题就将是如何成功地描述和解释这两个层次。

镜像(不很恰当的术语)

有机体对于其周围世界及自身的全部知识(1960:17-18)。

规划

操作序列对有机体内任何组织层次产生作用的加工过程(1960:16)。

执行

当规划正在执行控制操作序列时,有机体也就正在执行这个规划(1960:17)。

TOTE 单元

"TOTE"是"测试(Test)"、"操作(Operation)"、"测试(Test)"、"输出(Exit)"的首字母缩写,MGP 认为可以用这个分析单元代替 S-R 理论的反射弧。他们最初是在神经层次提到它的:"神经机制中发生的反射行为,并不能用简单的反射弧或者刺激一反射联结链条解释说明。牵涉更多复杂监控、测试的反射行为的解释要比反射弧解释能够更充分地满足各种条件"(1960:25)。"因此,反射行为的普遍样式是根据有机体内固有的某些标准测试输入能量,如果有机体的反应与测试的结果不一致,那么继续对其作出反应,直至差异消除才结束反射"(1960:26)。 51

在 TOTE 单元中的箭头(见下文),既可以表征神经冲动,也可以表征更高层次的加工过程,"反射仅仅被看作是实现 TOTE 样式众多可能中的一种,接下来的工作需要对 TOTE 单元进行概括,使它能够适用于绝大多数——我们希望的——对于行为的描述"(1960:27)。MGP 详细讨论了两种加工过程,即传送信息和传送控制——告诉系统接下来做什么。在标准数字计算机中,这两种过程均由程序完成,控制通过从一个指令到另一个指令的转换,而告诉机器接下来做什么,执行什么操作,以及在哪里存储结果。在 MGP 看来,"TOTE 单元……能够对普遍行为进行解释"(1960:29)。在解释复杂的具有层次等级的行为时,他们提出一个 TOTE 单元能够嵌套在另一个 TOTE 单元中。规划使复杂 TOTE 的层级结构得以概念化;行为可用包含在 TOTE 单元内执行操作的术语进行解释;"镜像"(有关世界的知识)提供条件,当需要时就会被测试。对这样一个复杂 TOTE 单元进行解释的例子是把钉子敲进墙内,如图 2.2 所示。

从 TOTE 层次结构到计算机的转变,是一种迅速且必然发生的趋势:"规划对之于有机体等同于程序对之于计算机"(1960:16)。恰好 TOTE 还满足如下两种功能:表征周围环境的面貌和控制系统接下来做什么。MGP 的镜像以及 TOTE 单元的测试,都包含着对环境的表征。GMP 的规划以及 TOTE 单元的操作,控制系统下一步要做什么。类似地,计算机的数据结构表征环境,程序是特定的控制转换——机器接下来做什么。而且,TOTE 单元嵌套于 TOTE 单元也类似于程序嵌套于程序。

52

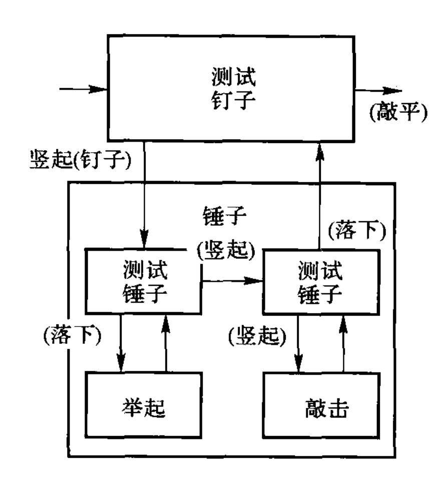


图 2.2　锤钉 TOTE 单元(源自:Miller, Galanter, and Pribram, 1960:36, 图 5;经作者许可复制)

到了 20 世纪 60 年代中期,"信息加工"研究取向已经普遍被人接受,Neisser(1967)是有关它的第一本教科书——这本书被普遍认为在这一领域里产生了重要影响。不久,林德赛(Lindsay)和诺曼(Norman)出版了《人类信息加工:心理学导论》,更加坚定了这一研究取向,并主导了心理学未来 10 到 15 年的发展。

注释

[1] 该书的销售情况并不乐观,4 年内只售出 80 本。

【思考题】

"经典条件反射"的要素是什么?
桑代克"效果律"蕴含的主要观点是什么?
桑代克"练习律"蕴含的主要观点是什么?
在华生的论述中,行为主义的四个主要议题是什么?
"斯金纳箱"具有怎样的结构?
什么是"操作行为"?
在行为主义后期,出现了哪四种趋势?
拉什利反对行为主义的主要观点是什么?
拉什利如何理解神经系统?
53 乔姆斯基对斯金纳的主要批评是什么?
拉什利和乔姆斯基对斯金纳的批评有哪些共同点?
"认知"传统是从哪两个领域(心理学之外)出现的?

信息的通用概念与技术概念分别是什么？
如何测量信息（技术概念）？
米勒发现的关于短时记忆的限制是什么？
米勒、加兰特尔和普利布拉姆提出的“镜像”、“规划”和“TOTE 单元”分别是什么？
TOTE 单元与计算机有哪些共同点？

【推荐读物】

概论

Gardner(1985)第 5 章强调了认知在心理学中的作用以及对认知科学发展的贡献。Leahey(1992)是近期较有影响的心理学史概论专著——该书的第三和第四部分尤其与我们关注的论题相关。Hilgard(1987)与本书有一些相同的材料，重点关注美国心理学，还包含更多的传记信息。Hernstein and Boring(1966)除了大量地从心理学史（以及哲学和心理学）中搜集名著外，在一些相关论题上并没有取得多少实质性进展。Pavlov(1927)是巴甫洛夫的经典著作。

行为主义

Skinner(1976)是一本权威性的关于行为主义的导论性著作。Fancher(1979)第 8 章也对行为主义进行了讨论。Boring(1957)的“行为主义”一节虽然有些过时，对于我们理解行为主义仍有帮助，且可读性较强。Marx and Hillix(1963)是一本经典著作，第 7 章“行为主义”和第 10 章“S-R 理论的种类”，从理论视角对行为主义的一些观点进行了概述，不过略显陈旧。对行为主义最新的讨论，尤其是有关斯金纳的贡献，见 Flanagan(1991)第 4 章。Hunt(1993)第 9 章对行为主义提供了具有可读性的概览。Smith(1986)对行为主义与逻辑实证主义的关系进行了研究。Dennett(1995)对自然选择在认知和行为主义的特殊应用进行了讨论。

其他流派活动

Flanagan(1991)讨论了这段时间内其他各种心理学流派的观点，其中第7 章 54 讨论了弗洛伊德的精神分析，第 10 章讨论了格式塔运动。

认知主义

除了本书中引用的那些著作外，在前面推荐的读物中也对认知主义有所涉及，如 Gardner(1985)第 5 章中后面的部分，Flanagan(1991)第 6 章以及 Hunt(1993)第 6 章。

3 生物学背景

3.1 引　言

55　神经科学从创立到现在，在很大程度上将心智看作脑**结构**的**功能**，一直存在**定位论**和**整体论**之争。神经科学首先确定了脑在思维（广义）和行为中的作用，然后解决了思维由脑质产生还是由脑质中的脑孔（脑室）产生的问题。接下来，按照功能和解剖学特征对脑进行了粗略的划分，例如划分了小脑及脑皮层。然后又进一步按照功能差异对脑皮层进行了细分，并发现了神经系统的基本结构。最后又发现了神经细胞的基本功能以及它们的细微结构。在很大程度上，从前一阶段到下一阶段的进步，有效技术的发展起到了决定性的作用，例如显微分辨和染色体技术。

3.2 脑室与脑基质

最早发现的是脑在思维（广义）和行为中的作用。古希腊人在心智产生自何处（“灵魂的居所”）的问题上有很多争论。很多人一直相信心脏是心理功能产生的最初原因，如因“四根说”（土、气、火和水）而广为人知的恩培多克勒（Empedocles，公元前 493—前 430）就持有这种心脏中心论。亚里士多德（Aristotle，公元前 384—前 322）同样认为心灵由心脏产生，但做了精细改
56　进，认为人脑的功能是心脏“火热和沸腾”的冷却器。而另外一些人则认为，大脑是知觉和思维的中心，如阿那克萨哥拉（Anaxagoras，公元前 500—前

428)。提出物质“原子论”的著名哲学家德谟克利特(Democritus,公元前460—前370)则相信一种三位一体的心灵观,认为:灵魂的第一部分位于脑,负责理智的事情,是不朽的;第二部分在心脏,与人的情感有关;第三部分在内脏,与人的欲望、贪念、需求以及“较低(水平)的”热情相联系,后两部分与第一部分不同,并不是不朽的。

接下来需要解决的问题是,思维是由脑质产生还是由脑质中的脑孔(脑室)产生。认为脑在人的思维中具有重要作用的人中,一部分人认为这种功能产生于脑中大的穴腔(脑室),而另一部分人则认为产生于其周围的脑组织。盖伦(Galen,130—200)提出过很多产生过重要影响的思想,其中就有这样一种观点:生命精气由左心室产生,经由颈动脉传送到大脑,脑中存在(由小动脉和小静脉交叉组成的)**异网**(一种神奇的滤网),在那里生命精气转化成更高级的精气。这些精气储存在脑室中,当需要的时候它们就会通过中空的神经使肌肉产生运动以及调节感知觉(但他并没有明确说明如何实现)。值得注意的是,盖伦将心智分为三个部分:想象、认知和记忆,它们与脑质(脑髓)相联结,但其功能却并非产生于脑质。到了公元4—5世纪,人们逐渐接受了这种心灵**脑室定位**的观点。如埃摩萨(叙利亚)主教尼梅修斯(Nemesius,390),认为感知觉位于两侧脑室,认知产生于中脑室,记忆则位于后脑室。这种观点以这种或其他形式几乎持续了上千年。

笛卡尔

他是自亚里士多德之后第一个提出一种新心理学的人。

勒内·笛卡尔(René Descartes,1596—1650)涉及了有关认知科学基础的几乎所有问题。他一出生就从母亲那里感染了肺结核(他母亲因此病逝),自小体弱多病,成年时身体也相对孱弱。他在耶稣会士学校里接受了数学和哲学教育(他们准许他可以直到中午才起床,他的这种习惯保持了一生)。他还在巴黎住过一段时间,并在军队服役。32岁时迁居荷兰,在那里完成了他一生中最重要的一些著作。后来,笛卡尔应邀到瑞典为女王克里斯蒂娜教授哲学,因清晨5点赶路遇到大雪而患感冒,因肺炎去世,时年54岁。 57

人与动物

笛卡尔认为,动物并没有心灵,它们的行为只用机械原理就可以得到解释,特别是应用反射原理。而人类同时拥有躯体和心灵,如果人的一个行为

是在无意中完成的，那么用机械原理也可以得到完全解释，与动物一致。但如果行为是随意的，那么就涉及人的心灵了。笛卡尔的这种身体机械观是他普遍机械决定论（不包括人类）世界观的一部分，据说这是笛卡尔在一次游览位于巴黎近郊圣杰曼-昂-雷法国皇家园林时得到的灵感。那里有一个“机械人”，由一些水管和阀组构成，游览者只要脚踩踏板，机械人就可以活动，甚至能够以现实的方式“谈话”。由此，笛卡尔想到机械人与神经系统的一种类比（这种观点流行于他那个时代）：

水管∷神经

水∷“动物精气”[1]

弹簧和发条∷肌肉

在生物学领域，笛卡尔可能是最早提出**反射弧**概念的人。他认为当外物刺激感官，就会拉动感官下面神经管中的细线，从而打开阀门，使脑室中的动物精气沿着神经管流到肌肉，触发肌肉运动（但他并没有说明如何实现）。动物精气的流动同样影响着部分心理功能，如感觉印象、情绪（爱悦、憎恨、欲望、欢乐、悲伤、惊奇）以及记忆（重复的经验会使脑内的一些孔道增大，因而动物精气更易流通）。但最终他的解剖学为哈维的解剖学所取代，亦如他的物理学为牛顿物理学所取代。

心灵与身体

笛卡尔认为，人的随意行为可以用意志得到解释：“意志的本性是自由，它从不会被约束……当它需求某些东西产生时，便与一种小腺体紧密结合在一起，以某种方式作用身体产生符合需求的行为结果，灵魂的整个行为便是如此”（《灵魂的激情》）。其中小腺体是指松果体，悬挂在两个前脑室之间，受到储存在那里的动物精气的影响，也会反过来影响动物精气。因此，笛卡尔似乎持有一种改进的脑室论（见图 3.1）。

58 在笛卡尔看来，心灵是明显不同于身体的东西，两者适用的原理也不相同[2]。他提出一种三段论支持这种观点，便是著名的“我思故我在”：

(1)我不能怀疑我作为一种思维物而存在（我思考，因此我存在）。

(2)我可以怀疑我作为一种身体而存在（我可能被一个邪恶的恶魔愚弄，它让我感觉这个身体是我自己的）。

(3)因此，我的心灵与我的身体并不相同。

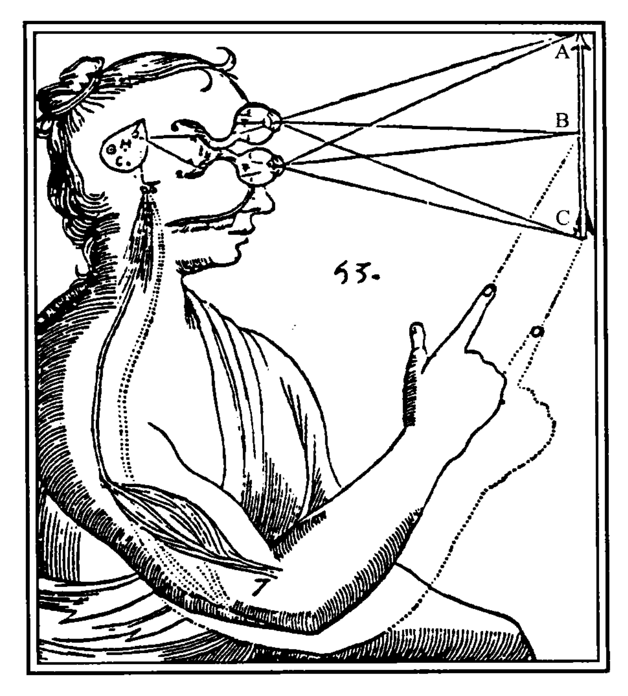


图 3.1　笛卡尔理论(来源:Finger, 1994:26,图 2.16;经 Oxford University Press 许可复制)

这个论点受到了许多中肯的批评,一般都认为它是有缺陷的;鲍勃·迪伦[①]和罗伯特·齐默曼是同一个人,然而人们可以怀疑罗伯特·齐默曼是否富有,却不怀疑鲍勃·迪伦。笛卡尔还基于下面的事实证明心灵与身体是不同的:

(4)身体不能进行思维,其本质具有空间广延性;心灵的本质是思维,不占据空间。

59

(5)身体可分为不同的组成部分;心灵不可拆分,它没有零件。

(对这两点的评论留作思考)所以**心灵**具有思维、非广延和不可拆分的特征,并且其意志行为是自由的,非决定的;相反,**身体**没有思维能力,是广延的,可拆分的,它的行为具有机械决定性。从笛卡尔的论述中可以推导出三个有关认知科学的重要结论。首先,人类心灵的内容完全是有意识的、内

① 鲍勃·迪伦(Bob Dylan)是美国具有重要影响力的唱作人、民谣歌手、音乐家和诗人,他原来的名字是罗伯特·齐默曼(Robert Zimmerman)。——译者注

省的，“就心灵具有思维的特征而言，如果它没有意识到什么的话，那么在心灵之中也就不会有任何东西……对于没有意识到的东西，我们不能进行任何思考”；其次，内省得到的内容具有“独断性”——任何其他人都不可能比你的思维结果更具有私有性；第三，这些内容只能内在于心灵之中，不仅独立于物理躯体，也与周围世界完全不同。

心一身交互论

如果物理躯体与心灵是完全不同的两种基质，那么它们具有怎样的关系呢？在笛卡尔看来，心灵能够摆动松果体，从而引起肌肉收缩。同样地，刺激作用感官产生的某种效果，也会作用松果体影响心灵。简言之，笛卡尔持有一种心一身因果交互论（在第 8 章我们将会回到这种理论）。

人

笛卡尔有时只用心灵来定义人：“我是谁？是一种能够思维的事物，那么思维又是什么呢？它是这样的东西，即能够怀疑、理解、确信、否定、决心、拒绝或者是能够想象和感觉”（《第一沉思录》），“这便是我（就是说，是我的灵魂决定了我是谁），与我的身体完全且绝对不同，可以脱离身体而独立存在”（《第一沉思录》）。今天自然科学（如生物学、神经生理学）的研究方法已经应用到认知领域，笛卡尔的观点自然与现代认知科学的成果相违背。但另一方面，他的某些观点与现代人们的看法也很相近：“我不仅像引航员生活在船里一样，固定寄居在我的身体里，而且我已经紧密地与它结合。就是说我已经与身体浑成一体，似乎已经组成了一个整体”（《第一沉思录》）。在笛卡尔最后一部著作《灵魂的激情》（1649）里，将人的行为划分为三种：一种属于心灵（理智的和意志的行为）；一种属于身体（生理的行为）；还有一种属于心灵和身体的“结合体”（情绪和感知的行为）。

60

3.3 皮质层定位论与整体论

高尔和施普尔茨海姆

第一个明确提出皮质层定位论的是高尔（Franz Joseph Gall，1758—1828），尽管这一问题在他之前就受到许多人的关注，这一理论的提出受到

许多因素的影响。高尔是德国人，从1796年开始在维也纳宣传他的理论，受到他的学生和后来的合作者施普尔茨海姆(Spurzheim，1776—1832)的追捧。1802年，政府命令他停止宣讲(迫于教会的压力，因为教会反对他的这种"唯物主义"学说)。在德国旅行一段时间之后，他于1807年来到巴黎。他们的第一部专题著作出现在1810年到1819年间，分为四卷本(第1卷和第2卷与斯普尔茨海姆合著)，标题为《从解剖学和生理学看神经系统的总体特征和脑的特定特征……》。高尔后来(1825)又完成了一部六卷本的著作《脑及其每个部分的功能》，10年后被译成英文。尽管受到大多数科学家的蔑视(有人说它是"人类愚蠢的总汇")，但颅相学为高尔赢得了广泛的声誉，并使他过上了宽裕的生活。

在我们讨论高尔的"颅相学"(通过颅骨隆起的外在表现确定心理功能的物理定位，这个名词最初由他的一个学生提出，但高尔并没有使用)之前，首先要认识到高尔之所以能够产生较大的影响，部分地与其熟练的医学技术相关，"所有人都会赞同他是一位杰出的脑解剖学家"，做了许多神经科学的基础性工作，包括脑体积的比较研究，发现皮层的大小一般与智能生物体的聪明程度有关。"在此之前从没有人如此清晰地表明，脑的体积与心理发展的平行关系"(Fancher，1979:45)。然而，高尔的这些成就却都被他令人生疑的推论所淹没：

(1)心灵可以分解为几种**具体官能**或者能力(高尔认为这些能力都是内在的)。

(2)不同的心理官能对应着脑内不同的**具体部位**。

(3)不同的具体心理官能所处的物理部位，可以通过**脑质块的大小**显现出来。

(4)质块的体积可**在颅骨表面得到反映**，这样通过头骨学(颅骨测量术)的知识就可以知道这些能力的相关情况。 61

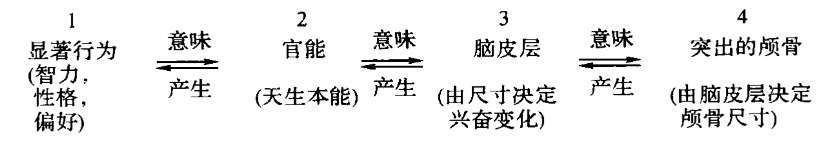


图3.2 高尔的解释(来源：Young，1970:36)

高尔的方法如图3.2所示，通过观察1和4，便可推知2和3。高尔曾说过，他第一次产生这种颅相学的想法是在9岁时，他观察同学中眼睛较大的

人通常语言记忆也较好。他由此猜测负责语言记忆的那部分脑区比普通人要大,导致把眼睛挤压出来。高尔写道:"给我留下很深印象的是这两种完全不同的情形总是同时出现,我可不认为是出于偶然,现在更加确信这一点。从那时起,我就开始怀疑眼睛的一些特征与记忆能力之间必定存在某些关联"(1835,vol.I: 57-8)。高尔并没有说他是怎样"更加确信"这种关联的,很可能出于半虚构半真实的观察。无论如何,他的这种方法注定要被抛弃。举个例子,他得出人的**破坏天性**位于耳朵上面(在高尔看来,脑的两半球是一样的),原因是:(1)这是颅骨上最宽大的一块头盖骨;(2)曾有一个学生的头盖骨在这个位置上突显,"他非常喜欢折磨动物,后来成了一名外科医生";(3)一名药剂师这个部位的头盖骨非常发达,后来成了一名行刑人。他还用半真半假的态度检验他的理论,高尔这样没有原则、没有依据地对心理官能相对应的脑部位进行选择的做法遭受众多质疑,不同的人很有可能会得出不同的结果。高尔最终确定了27种心理官能,其中19种同样发生在动物身上(高尔的这个列表主要受到苏格兰哲学家托马斯·里德(Thomas Reid)的影响,里德曾粗略地提出过存在24种心理活动功能以及约6种与智力水平相关的能力),如下(Corsi,1991:155):

1. 繁殖后代的本能
2. 对子孙的爱
3. 依恋和友谊
4. 保护自己和个人财产的本能
5. 残虐、破坏
6. 聪明、敏锐、多智
7. 自制力和倾向稳定
8. 骄傲、傲慢、强悍、热衷权威
9. 虚荣、野心、热爱荣誉
10. 细心谨慎、深谋远虑
11. 对事情、事件的记忆能力
12. 空间关系的知觉
13. 回忆认识过的人的能力
14. 词语、姓名和语义的理解力
15. 表达能力和语言天赋
16. 颜色辨别或者绘画天分
17. 判断音调关系或者音乐天赋

18. 数字关系的判断力
19. 机械、构造和建筑的能力
20. 擅于比较
21. 思想深度和纯粹哲学精神
22. 幽默感和讽刺能力
23. 诗歌天分
24. 善良、仁慈、可爱、同情、灵敏、道德感
25. 模仿、模拟的才能
26. 上帝和信仰的崇拜
27. 坚定不移、忠贞不渝、不屈不挠、坚韧不拔

高尔的合作者施普尔茨海姆给颅相学带来了现代特征，他采用了"颅相学"这个名词，并且区分了理智和情感，使高尔提出的概念更加精细化，还增加了官能的数目（见图3.3和图3.4）。施普尔茨海姆和高尔在1813年彼此分开——高尔继续留在巴黎，而施普尔茨海姆则到美国讲学（现在还可以在哈佛医学院沃伦博物馆看到他的颅骨）。

弗楼伦

对颅相学最著名的反对者是皮埃尔·弗楼伦（Marie-Jean-Pierre Flourens，1794—1867）。他是一位受人尊敬的解剖学家、心理学家和神童，在19岁时就发表了他的第一篇学术论文并获得医学学位。弗楼伦推崇实验室实验研究，方法是这样的：先切除小动物（鸟、兔子、狗）脑的一部分（切除手术），就是颅相学中认为与具体行为特质相联系的区域，然后照料动物直到它们伤口愈合并恢复健康，然后严格控制实验条件，比较做过切除手术和没有做过切除手术动物的行为。由于经常得不到颅相学所预测的结果，弗楼伦开始提出他的机能统一论："所有的感觉、知觉和意志行为都同时发生在脑器官的同一位置上，感觉、知觉和意志本质上是同一种机能"（Finger，1994：36）。但是由于他在给脑做切片的时候，切片横跨了解剖上和功能上完全不同的脑区，所以这种推论并没有完全得到证明。随后的研究表明，弗楼伦采用了正确的方法，却得出了错误的结论；而高尔作出了正确的结论，却使用了错误的方法。

63

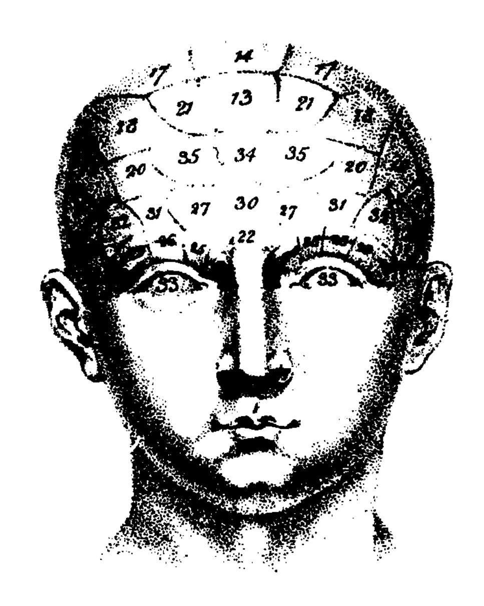
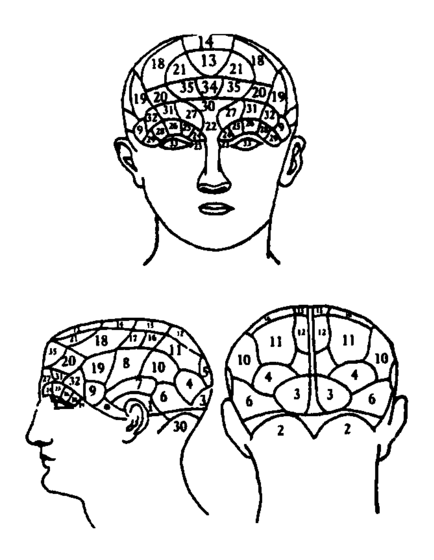


图 3.3 施普尔茨海姆 1825 年插图(来源:Finger, 1994:33,图 3.2;经 Oxford University Press 许可复制)

布洛卡

为脑机能定位论的合理发展做出历史性重要贡献的是前面提到过的法国医生和科学家保罗·布洛卡(Paul Broca,1824—1880)。早在布洛卡的研究工作之前,因中风(常常因血凝块导致脑供血受阻)而引起的**运动性失语症**(不能用语言进行交流)就早已引起人们的关注。具有讽刺意味的是,高尔也作过同样的解释,"最早明确记录了语言缺陷与脑皮层左额叶创伤的关系"。一般认为,第一个被发现因此区域损伤而引发失语症的病人,是 1861 年布洛卡的一个患者莱沃尔涅先生(Monsieur Leborgne,人们当时只称呼他为"Tan"先生)。布洛卡那时发表了现在人们认为是"历史上有关脑皮层机能定位论最重要的临床论文"(Finger,1994:38)。Tan 先生的故事也很有趣。与布洛卡同时代的名叫奥贝坦(Aubertin,1825—1893)的医生,比布洛卡更早对语言的区域定位学说产生了兴趣,他对一位特殊的病人进行了长

64



“心灵官能与其对应官位”，引自施普尔茨海姆《颅相学或精神现象原理》(1834)

情感官能		智力官能	
倾向	性情	感知	思维
? 生的欲望	10 谨慎	22 个体	34 类比
* 食物需求	11 被人认可	23 结构	35 因果逻辑
1 破坏	12 自尊	24 体积	
2 性爱	13 善良	25 重量和抗性	
3 爱子	14 敬畏	26 颜色	
4 眷恋	15 固执	27 位置	
5 恋土	16 认真	28 次序	
6 好斗	17 希望	29 计算	
7 私密性	18 惊异	30 不确定事项	
8 好奇	19 理想	31 时间	
9 积极	20 高兴	32 声音	
	21 伪装掩饰	33 语言	

图 3.4　施普尔茨海姆 1834 年插图(来源:Boring, 1951:55,图 1;经 Prentice-Hall Inc. 许可复制)

时间研究后，向机能定位论持怀疑态度的巴黎人类学协会(Society of Anthropology in Paris)发出挑战，说:“如果给病人做尸体解剖，发现额叶是完好的，那么我就放弃我所坚持的观点”。就在这件事发生之前，布洛卡接收了一位 51 岁的病人，他在 30 岁时就得了失语症。因为他能做的事就是发出类似“Tan”的声音，于是“Tan”便成了他的绰号，尽管极少(在一次与布洛卡见面时)但偶尔也会说出:“Sacre Nom de Dieu!”布洛卡让奥贝坦为 Tan 做了检查，之后奥贝坦指出:“毫无疑问，这个病人皮质层的左额叶受到了损伤。”几天之后，病人死于坏疽，布洛卡做了尸体解剖，并把他的大脑带到了

65

下一届的人类学协会会议上。大脑左侧一块鸡蛋大小的区域已经受损，损伤区的中心(假定最初损坏的点)恰好处于左额叶第三回沟的下半部。

几个月后，布洛卡有了另一个病例，一位84岁突然不能讲话的老年人，在这位老年人去世后，人们在进行尸体解剖时发现，恰恰是在Tan损伤区域中心的位置有一块小的损伤。布洛卡又收集了其他几个病例都支持他得出相同的结论。布洛卡记录道，每一个病例损伤部位都在大脑**左半球的同一位置**——损伤右半球的同一位置(病人)并不丧失相同的能力。为了纪念布洛卡，现在这块区域便称之为“布洛卡区”，这种症状称为“布洛卡失语症”。下面这段对话是“布洛卡失语症”的典型症状(Akmajian et al., 517)：

医生：告诉我，在退休之前你是做什么的？

失语症患者：Uh, uh, uh puh, par, partender, no.

医生：木匠？

失语症患者：(点头表示是)Carpenter, tuh, tuh, tenty [20] year.

医生：告诉我这张照片上都有什么？

失语症患者：Boy... cook... cookie... took... cookie.

维尔尼克

1870年，卡尔·维尔尼克(Carl Wernicke, 1848—1905)又前进了一步，表明如果损伤的是颞叶的一部分会导致语言混乱，特征是缺乏对语言的理解而不是不能讲话。在维尔尼克时代，人们对大脑心理功能的理解是这样的，不同类别的感觉信息投射到脑皮层感觉功能区的不同点上，以“图像”的形式储存在它周围的组织中；具体的行为动作，也以“图像”的形式存储在运动功能区；还有一部分脑组织，认为是连接感觉和运动区中心的**联合区**。维尔尼克是第一个为运动性失语症(布洛卡失语症)作出解释的人，即认为是由于掌管储存语言清楚表达“图像”的区域受到了损坏。但他最著名的工作是发现并解释了(与布洛卡失语症)相反的另一种症状(现在称之为维尔尼克失语症)：病人可以非常流利地讲话，甚至对常见单词的发音从没出过错，但却不具有理解能力，这是由于损伤了靠近听觉区的颞叶部分而引起的失语症。病人可以听见别人讲话，知道有人在与他对话并试着做出反应，但就是不理解别人到底说了什么。如下面的例子：

医生：你为什么来医院？

病人：伙计，我有些发汗，有点紧张。偶尔会被抓住，但我并没有提

及 tarripoi(原话如此),一个月以前,我做了很多,然而另一方面,你该知道我的意思,我不得不到处转悠,检查,trebbin(原话如此)和所有类型的职员一样。(Fancher,1979:69)

机能定位论是基于外伤病例而提出的,不久都在实验中得到了验证。特别是精神病医生爱德华·西齐格(Eduard Hitzig,1839—1907)和解剖学家古斯塔夫·弗里奇(Gustav Fritsch,1838—1927)所做的工作。1870 年他们在柏林西齐格家里的梳妆台上,用细微电流直接刺激狗的大脑,发现刺激不同的皮层位置会引起狗做出不同的动作。大卫·费里尔(David Ferrier,1843—1928)重复了弗里奇和西齐格的实验,在猴子身上发现了许多更加精细和复杂的动作,如眼睑抽颤、单指运动。从此,解剖学家们开始逐渐揭示脑皮层的精细结构,特别是科比尼安·布罗德曼(Korbinian Brodmann,1868—1918)还制作了一张具有 52 种不同功能区域的脑皮层分布图(见图 3.5)。然而还没有一个所有人都认可的标准辨别这些区域,不同的研究者经常得出不同的分布图。很明显,神经中枢系统的提出(此时已经出现了神经功能的突触学说)和反射弧理论为这种观点提供了非常坚固的基础,即心理活动产生于原始观念的联结,由大脑中特殊的连接通路和中枢实现。

67



图 3.5 1909 年布罗德曼脑区图(来源:Finger, 1994:42,图 3.15;经 Oxford University Press 许可复制)

3.4 神经网状理论与神经元原理

英国伦敦人罗伯特·胡克(Robert Hooke)于1665年发明了显微镜。1718年,荷兰代尔夫特人安东·冯·列文虎克(Anton von Leeuwenhoek)发表文章说,利用显微镜看到了牛的神经末梢(时年84岁),声称发现的是很多单个的空心管,就是类似于笛卡尔(或者其他人)曾想象的那种从感觉器官通向大脑(感知),或者从脊髓通向肌肉(行为)的输送"动物精气"的那种空心管。问题是他使用的早期显微镜会随镜片的不同而反射不同波长的光,也就是"色差现象"。考虑到神经物质处于极小的数值范围内,所以这是一个很严重的问题。直到1820年,出现了功能更强和没有色差干扰的显微镜之后,才允许人们使用"显微镜"观察、确认神经细胞的胞体及其外延部分。第一个相当精确地看到神经细胞和神经纤维显微结构的是被称为"历史的发现者"的浦肯野(Jan Purkyne,1787—1869)。浦肯野最先学习哲学,后来还发现了"浦肯野位移"①。1836年浦肯野的学生加布里·瓦伦丁(Gabriel Valentin,1810—1883)绘制了一张神经细胞(a "kugeln" or globule)的显微图像,这是"第一份生物学中有关神经细胞的绘制图"。瓦伦丁清晰地描绘了胞膜的轮廓,并记录了神经细胞常常带有一个尾巴状的附属物,但他认为胞体和这个附属物都单独地被一种皮鞘包裹着,"因此,瓦伦丁与神经纤维发端于胞体这一发现失之交臂"。1837年,浦肯野在布拉格做了一次演讲,描述了小脑皮层上的一种大的神经节细胞,现在称为"浦肯野细胞"。浦肯野还注意到,"尾状末端的方向朝外,以两种**组织派生物**的形式消失在灰质中",那种组织派生物后来被称为"树突棘"。但他也没有发现这些纤维与胞体的联系。1838年,西奥多·施旺(Theodor Schwann,1810—1882)提出了"细胞理论":整个胞体,包括所有细胞内部和外部的部分,组成一种单个细胞。细胞理论适用每种器官组织,除了神经系统,主要原因是存在下面两个困难:

68

1. 显微镜并不能说明是所有的神经纤维均直接由神经细胞产生,还是部分神经纤维可能独立存在。
2. 并不能找到长且细的树突有**明确的末梢**,也不能确定它们是否

① "浦肯野位移"是指在不同的适应状态下对有色光的视觉灵敏度不同的现象。在明适应时对红色和橙色看起来较亮,而在暗适应时则对蓝色光看起来较亮。——译者注

与邻近的细的树突联结在一起，形成一种**连续的神经网络**。

第一种观点和另一种观点即神经元之间在接触点处相互作用而发生联结，便形成了“神经元原理”。第二种观点假定活性以连续的方式在整个树突网中扩散。1833 年，埃伦贝尔(Christian Ehrenberg)将神经纤维与血管系统的毛细血管作了比较，血液循环的动一静脉系统的连贯性为后来的“神经网络理论”提供了模型[3]。同时，研究者们也在尝试着是否可以减少这种种类繁多的树状模式，最后将所有神经纤维分为两类，即我们今天所说的“树突”和“轴突”(卡哈尔的“动力极化原则(Law of Dynamic Polarization)”表明，每个神经元都可以作为一个整体——树突接收信息，轴突输出信息，后面阐述)。戴特斯(Deiters)和柯力克(Koelliker)分别于 1865 年和 1867 年提出轴突是相互独立的，而树突则可能联结成网络。

根据柯力克主要于 1853 年所完成的工作进行判断，他也许是第一个在事实上确立了神经纤维产生于神经细胞的人。1865 年，戴特斯第一次指出了我们现在称为树突和轴突的区别，“戴特斯通过对树突和单个轴突的仔细观察，使他直接走上了神经元原理的道路，甚至已经通向了现代。然而讽刺的是，在介绍他的第二组精细纤维时，却为网状(神经网络)理论作出了贡献”(Shepherd，1991:47)。需要新的和更先进的技术才能解决这个难题，这把前进的钥匙便是 1870 年高尔基发明的染色法。

高尔基

高尔基(Camillio Golgi，1843—1926)在 1873 年获得了一个重要的发现——水银染色法，这是在米兰附近他工作的一家医院厨房里发现的。这个方法很难可靠使用，但一旦成功，神经细胞就会在黄色的背景下被清楚地染成黑色，从而相关的全部显微特征都能够显现出来。 69

高尔基认为，轴突形成一个稠密的相互融合的网络，是一个循环系统(“交叉融合(anastomosis)”)，完全可以通过染色技术得到的结果证实。然而，“事实上，这种方法并不足以得出这种结论，因为微细的树突末端并没有被完全染色，它们实在太微细了，所以无论如何也不能依靠光学显微镜清楚地证明它们之间‘交叉融合’，高尔基的神经组织网络理论忽略了这个致命的误解”(Shepherd，1991:91)。高尔基还认为，树突的主要作用是提供营养。这种神经网络的观念使他反对脑定位理论而支持脑功能整体论。1883 年，他写了一篇很长的文章，对他的研究发现作了回顾和总结，其中尤其与网络观念相关的内容如下：

> 十一:在中央神经器官的所有灰质层中,存在一种精细而复杂的扩散神经网络……可以被分解为微长细丝,(这些细丝)逐渐形成一个交织的网络,因此失去了它们特有的独立性……这里描述的网络,显然使中央灰质广阔区域内的细胞分子构成了一种解剖上和功能上的整体。
>
> 十五:据前面所讲的,可以推出另一个必然结论,即所谓的脑机能定位理论,严格来说……并不能被精微的解剖研究结果所支持。(Shepherd,1991:99-100)

令人惊讶的是,高尔基在进行他"卓越"的研究生涯的同时,还兼任帕维亚大学的校长和罗马参议员。

卡哈尔

在此时期有很多解剖学家反对(神经)网状理论,以圣地亚哥·拉蒙·70 卡哈尔(Santiago Ramonyy Cajal,1852—1934)最为强烈。1887 年,卡哈尔在马德里观察了高尔基的染色样本后,通过切制厚截片研究神经元前端的髓鞘而改进了染色方法。1888 年,卡哈尔在他的第一篇论文中提到,轴突和树突都找不到"交叉融合"形成连续网络的证据。随后的研究更加支持了他的这种判断。他于 1889 年发表文章提出,所有的神经细胞都是独立的元素。

1889 年,卡哈尔来到柏林,在一个讨论会上详细阐述了他的观点。在那里,他遇到了威廉姆·冯·沃尔德耶(Wilhelm von Waldeyer,1836—1921)。沃尔德耶深受卡哈尔的影响,在 1891 年写了一篇流传很广、颇具影响力的理论评论支持卡哈尔。卡哈尔认为,"神经元"(神经细胞)是神经系统解剖、生理、新陈代谢和遗传的基本单元。这里,他引入了"神经元"术语,提出了他的**神经元原理**:

> Ⅰ.所有神经纤维的轴索(轴突)……表明,它们直接发端于胞体。并不存在纤维网络或者任何与网络有关的东西。
>
> Ⅱ.所有的神经纤维都以"树状尾端"终结,并没有网状和交叉融合的形态。(Shepherd, 1991: 181-2)

1894 年,卡哈尔应伦敦皇家学会邀请在声名赫赫的克鲁尼安讲座(Croonian Lectures)中所作的报告(题目是"神经细胞的细微结构"[①]),成为他在 1909 年出版的有关神经系统神经结构的两卷本著作的基础。这是神经

① 原文为法文"La fine structure des centres nerveux"。——译者注

科学史上非常重要且关键的一次报告，总结了他之前的实验研究，以及高尔基、戴斯特和柯力克等人的工作，具有里程碑式的意义。他这样令人振奋地写道："探究神经系统中不同神经细胞的类型和它们的交互作用，这种理想比其他的研究来说，早已根植在了连续几代科学家的心中"（同上，254）。卡哈尔提出，每个神经细胞由三个部分组成，各部分具有完全不同的功能：胞体和"原生质的延长部分"（现在称"树突"）接受刺激，"轴索"（现在称"轴突"）负责传递刺激，轴索尾端负责发送刺激，基本上与现代的观点相一致。然而神经元理论的追随者们并不能解释：如果它们不是互相融合的，那么如何相互进行交流？

查理斯·谢灵顿（Charles Sherrington，1857—1952）推测神经元之间及神经元与肌肉之间存在着间隙连接点，并在1897年称这些连接点为"突触"（"synapse"取自希腊语，为了强调起"联结"作用）。他在1897年的一篇著名论文中写道："迄今为止，我们已经逐渐明确，树状枝末梢和与其相互作用的树突或胞体物质之间并不是连续的，只是互相联系而已。一个神经细胞与另一个神经细胞的这种特别的空间联合称为突触。"在此之前，埃米尔·杜·布瓦-雷蒙德（Emil du Bois-Reymond，1818—1898）在19世纪70年代曾说道，倘若兴奋从神经传递到效应器细胞的动力可以由电或者化学反应提供——那么具体的细节就要等到20世纪凭借技术的发展才能揭示出来。 71

迟至1906年（正是高尔基和卡哈尔因他们的研究工作共同获得诺贝尔奖的那一年），高尔基对神经元原理的三个中心论题表示反对。这三个论题分别是：(1)神经元是一种增殖单元；(2)神经元是一种细胞；(3)神经元是一种生理单元。高尔基依然坚持轴突互相融合交叉而形成一张大的整体统一的网络的观点。他在接受诺贝尔奖发表演讲时说道："关于神经元问题的这一说明，与其说是结论还不如说是对这些现象所作的一种整合，又把我带回了原点——也就是说，沃尔德耶用来支持神经元个体性和独立性的理论论据是经不起深究的。"因此，他也同样反对脑机能定位理论。

3.5 20世纪前半叶

拉什利

由于卡尔·拉什利（Karl Lashley，1890—1958）与他的同事，尤其是谢

菲尔德·弗兰兹(Shepherd Franz,1874—1933)的研究工作,使整体论(回忆弗楼伦的观点)又重新流行起来。在他们早期的研究中(1917),弗兰兹和拉什利发现受到不同程度损伤的(动物)大脑,仍然具有完整的明度识别和对简单迷宫的学习能力。拉什利因此得出智力水平的衰减随着损伤面积的大小而发生变化,与损伤的位置无关的结论(见图 3.6)。1929 年,拉什利对这项研究工作进行了总结,提出了他著名的等势原理和整体活动原理:

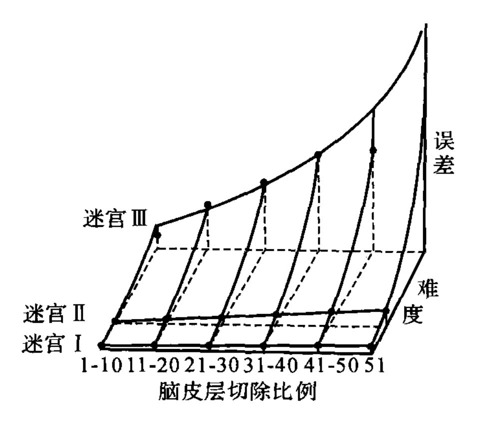


图 3.6 拉什利脑皮层切除实验结果(来源:Fancher, 1979:76:2-3;经 W. W. Norton & Co. Inc. 许可复制)

等势(Equiptentiality):"我使用'等势'这个术语,是为了说明脑皮层的任何完整部分都能执行明显的脑功能,但在效率上有所不同。只有当脑皮层全部被损坏时,其功能才将完全消失。只有对联合区与比之简单感知或运动协调在功能上更复杂的脑区,它们的效率和所涉及的功能特征才随不同的脑区而发生变化。"

72

整体活动(Mass action):"我以前已经给出了证据,现在的研究还将前进一步。等势原理并不是绝对的,它还要遵循整体活动原理。整体活动原理指,单一复杂功能的作业效率随着大脑损伤程度递减。在某一脑区内,它的某一部分并不比另一部分的功能更具特异化。"

我们注意到,拉什利原理的适用条件并不包括感觉和运动机能。20 世纪六七十年代,勒特文(Lettvin)等《蛙眼告知蛙脑什么》以及休贝尔(Hubel)和威舍尔(Wiesel)的《视觉的脑机制》等在单根纤维(和单细胞)方面的研究工作,对这两个系统提出了非常极端的定位理论。

赫 伯

最后,我们将介绍唐纳德·赫伯(Donald Hebb)在他《行为组织》(1949)中综合脑机能定位理论和整体论所做的一些工作。赫伯的研究从拉什利停止的地方开始——尝试寻找一种方式解释大脑在回忆一些事情时,同时具有发散和定位两种功能的困难。赫伯的观点深受神经解剖学家罗朗特·德·诺(Lorente de Nó)的影响,德·诺通过对神经线路的分析而认为脑中存在"折返通路(re-entrant)"式的神经环。赫伯对这一观点进行了概括,认为在脑中存在具有"反响"功能的神经回路,并且这些回路可以看作是一种简单的封闭环(见图 3.7)。 73

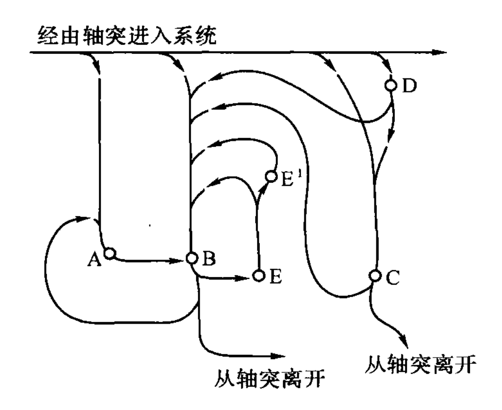
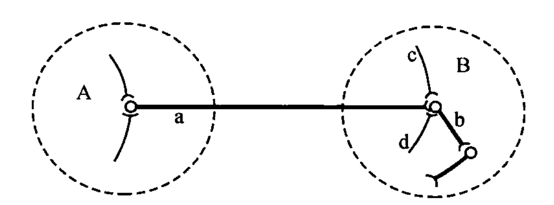


图 3.7　赫伯的神经环路图(来源:Hebb, 1972:70,图 23,图 3.1-3.10;经 W. B. Saunders Co. 许可复制)

在赫伯看来,行为模式建立在特异细胞(具有特殊功能的特定细胞)相互之间因长时程的联结作用而形成细胞集合的基础上。但这些神经回路是怎样产生的呢?赫伯认为,这是由于存在着一种神经机制——如果细胞因同时发生激活而相互联结,这便是所谓的"赫伯假设"。"赫伯假设"(并非赫伯本人提出)包含两种情形,均可在赫伯的著作中找到:首先,如赫伯所说,"当细胞 A 的一个轴突和细胞 B 很近,足以对它产生影响,并且持久地、不断地参与了对细胞 B 的兴奋。那么在这两个细胞或其中之一会发生某种生长过程或新陈代谢的变化,以致 A 的功能得到加强,如成为能使 B 产生兴奋的细胞之一"(1972:62)。第二种是指细胞 A 和细胞 B 同时发生激活的情形,如图 3.8 所示。



突触连接机制。A 是一组激活神经元，a 是 A 中明确标识出的一个神经元，B 是另一组同时激活的神经元，其中有两个神经元明确标识。既然 a 与 b 同时激活，并且 a 轴突与 b 接近，那么 a(为之激活更为有效)就将与 b 发生连接。

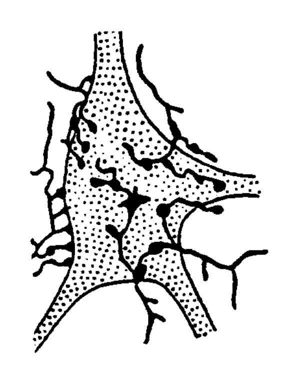
图 3.8　赫伯假设(来源：Hebb, 1972:64,图 29)

74　这种同时“同时发生激活”为联结主义在神经层次上提供了一种模型。随时间增加，细胞集合(整体论)会被固定在某一较大的**相位序列**上，使多个细胞集合加入到细胞集合的环路中，从而产生更为复杂的行为方式。

20 世纪中期神经元理论的发展

20 世纪中期，卡哈尔的神经元原理完全整合了神经元、突触和神经激活等概念，形成了新的神经元理论。下面将介绍这种神经元理论的一些基本内容(参见：Boring et al. ,1939；Hebb,1972)

神经元细胞具有多种不同类型(后面章节将具体阐述)，每种类型的神经元细胞都是**树突**接受信息，**轴突**发送信息。一般来说，一个神经元有很多树突但只有一根轴突，轴突的端点便是神经元的末端。**胞体**也可以直接接收信息(见图 3.9)。



突触：突触节点(黑色圆点标识)与胞体(灰色点画标识)建立联系。图中只画出了一小部分突触节点，实际上整个胞体和树突完全被突触节点所覆盖(参考 E. Gordner, *Fundamentals of Neurology*, W. B. Saunders)。

图 3.9　突触节点与细胞体的连接(来源：Hebb, 1972:68,图 28)

突触是轴突与胞体或者树突的联结点。轴突末端变大的部分称为**突触节点**。轴突(和胞体)以“**全一或一无**”的方式传递信息。它类似于“导火线”——在每一个点上把所有储存的燃料用光,这样在传递时就不会衰减。而树突则不同,它是衰减的:“[树突]就像弓箭,拉伸的幅度小,传递的距离也短”。且随着距离的增加,树突兴奋程度并不是以“全一或一无”的方式,而是**递级**发生变化。

神经冲动是神经系统中基本的信息传递过程,具有如下特征:

75

1. 冲动是一种沿神经元传递的**电的**和**化学的**变化。传递速率随神经纤维直径不同而发生变化:最小直径的传递速率约 1 米/秒,最大直径的传递速率约 120 米/秒。

2. 冲动通过突触引起下一神经元产生相似的冲动,或者引起肌肉收缩,或者使腺开始分泌。

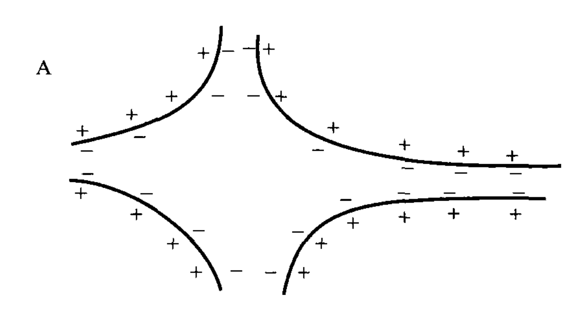
3. 传递一个冲动之后,神经元需要一定时间进行“充电”才能传递下一个冲动,这种**绝对不应期**约 1 毫秒。因为这种绝对不应期,一个神经元的最大兴奋率约为*1000* **次/秒**。

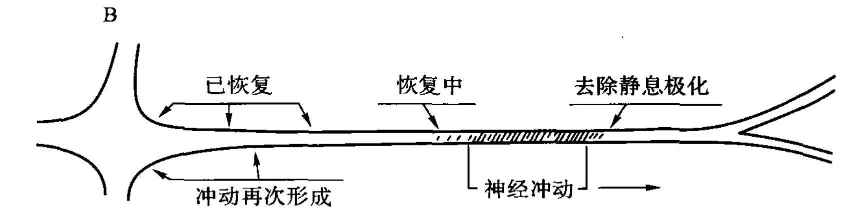
4. 神经元在激活后不能立即重新激活,但一段时间后却可以被一个更强烈的刺激激活。这就是**相对不应期**,持续时间为 0.1 秒。尽管这个强烈的刺激并不能产生更强的冲动,但缩短了上一个不应期的时间,增加了神经元的工作**频率**——将刺激的强度转换成激活的频率。细胞在长时程的高频率激活后,细胞因钠离子浓度的变化而产生疲劳,大约需要一个小时左右才能恢复正常水平。

5. 动作电位的产生,是由细胞外的**带正电荷的钠离子**通过半透膜进入细胞内,与细胞内的**钾离子**进行交换,从而增加了半透膜外侧的**负电荷**。接着轴突邻近部位也发生这种不稳定性,产生相同的过程。这种不稳定被打破之后立即会把钠离子从内侧泵出,恢复到原来的平衡状态(见图 3.10),这个过程需要大约 1 毫秒(在较大的神经纤维中需要 0.5 毫秒,小的则需要 2 毫秒)。

6. 细胞在激活后会产生**抑制**。这就是细胞的**超极化**与去极化,极化是指因半透膜两侧正负离子量的不同而引起的电位差。

7. 细胞能够对从不同来源的输入刺激进行**加和**,发生在细胞膜上或者在轴突中。因为单一输入引起神经元激活的概率较小,加和所有的输入就能增加激活的概率。





(A)静息状态下神经元的极化。(B)冲动(阴影区)沿轴突的传递:表明轴突可同时传递两个或更多冲动。只要第一个冲动已经移动到神经纤维中并且胞体已经复原,那么第二个冲动就可在“已恢复”区开始传递。整个过程事实上比图解A所示更为复杂。

图 3.10 神经极化与冲动传递(来源:Hebb, 1972:70,图 29)

注释

[1] “动物精气”指血液在输送到脑之前,被过滤出来的高度纯净的血液成分。

[2] 在源于古希腊尤其是亚里士多德的形而上学看来,实体是一种不需要依赖任何其他事物而独立存在的东西。

[3] 在某种程度上,当代的“联结主义”认知模型(见第三部分)可以看作是神经网络理论的后续发展,但只是针对功能层次而不是结构层次。

【思考题】

笛卡尔的心身关系理论的主要内容是什么?

什么是“颅相学”?

高尔检验颅相学使用的方法是什么?

高尔提出了多少种心理功能?

施普尔茨海姆提出了多少种心理功能?

在高尔和施普尔茨海姆的观点中,反映了心理功能定位论的典型示例有哪些?

对大脑语言功能“定位”最初的证据是什么(布洛卡、维尔尼克)?

什么是“布洛卡失语症”?

什么是“维尔尼克失语症”？ 77
神经网络理论(高尔基)和神经元原理(卡哈尔)之间的争论是什么？
为什么这个问题在很长时间内得不到解决？
沃尔德耶的贡献是什么？
什么是“神经元原理”？
从哪四种意义上来看，神经元是神经系统的基本单位？
等势原理(拉什利)的证据是什么？
什么是“赫伯假设”？并与詹姆斯假设作比较。
神经元的基础结构是什么？
什么是基本操作循环？
对神经元是神经系统的解剖单元观点的挑战是什么？
神经元是神经系统解剖单元这一观点的哪三个特征受到了挑战？

【推荐读物】

有关神经生理学历史

了解神经系统研究史的首选读物是 Finger(1994)。Corsi(1991)同样也包含一些较好的历史文本，对各种历史理论都作了相应的评论。在 Shepherd(1991)中，对神经元原理的兴起有很多吸引人的细节描述，其中第 239－253 页阐述了卡哈尔早期在克鲁尼安讲座产生的影响。本章所讲内容主要依靠的就是这些资料，不过有些细节经过了处理，使之更易于阅读。对于笛卡尔心理学更多 78 的讨论见 Hatfield(1992)及其参考文献。有关笛卡尔内在论的更多讨论，参见 Mcculloch(1995)第 1 章。

Young(1970)总结了大脑功能定位论的历史，其中有关高尔和布洛卡的讨论分别参见该书第 1 章和第 4 章。在 Boring(1957)第 3 章中，包含了关于高尔和施普茨海姆的一些传记和文献方面的有趣信息。关于高尔非颅相学的重要观点对心智结构的影响，在 Fodor(1983)中作了值得关注的讨论，同样也可参见 Corsi(1991)第 3 章和第 5 章。

有关生物学

对神经元和神经系统生物学较权威的导论，见 Shepaerd(1994)，尤其是第 1－9 章，同样也可参见 Kandel et al.(1995)。Beatty(1995)是一部导论教材，其中包含了许多与本书相同的论题并作了详细阐述。有关中世纪对于神经元的知识状态的权威而详尽的论述，见 Brink(1951)。在 Gardner(1985)第 9 章中，就有关神经科学对于认知科学的贡献，作了高度明晰的概括。在 Churchland(1986)中，对神经科学的发展历史作了简要回顾，并对当代神经科学理论和官能神经解剖学进行了概述，还讨论了有关哲学寓意。Stillings et al.(1995)第 7 章对神经科学作了很好的概述。

4 神经-逻辑背景

4.1 引　言

79　本章通过对神经元和神经工作网络(neural networks)的形式类比(或者理想状态)的研究,概述神经系统的计算特征及其发展过程。这种形式类比通常称为"神经网络(neural nets)"。我们首先介绍麦卡洛克和皮茨(McCulloch and Pitts,1943)关于神经系统开关逻辑(switching logic)所取得的开拓性成果。在他们研究工作的基础上,1960—1970 年间产生了两种不同的研究分支。其中一个分支,运用神经开关回路,精确说明了神经系统如何能被看作是一种二值计算的计算机。麦卡洛克和皮茨的这篇论文是冯·诺依曼在 1945 年 EDVAC 报告中唯一引用的文献,促进了心智数字计算理论的发展。因为麦卡洛克-皮茨网络没有学习能力,不得不需要使连接阈值与/或连接强度具有可变性,这就自然地引发了罗森布拉特(Rosenblatt,1958)关于感知器的研究。一个简单的感知器就是一种有着可变阈值与/或连接强度的双层麦卡洛克-皮茨单元网络。接下来介绍明斯基和帕佩尔特(Minsky and Papert,1969)对感知器的批评。感知器理论最终发展成为当代认知科学的联结主义。因此,麦卡洛克和皮茨的这篇论文被认为是认知科学史上的一个分支点,一方面通向序列的、数字的计算理论;另一方面通向平行的、分布式的加工理论。然后我们转向勒特文(Lettvin)、马图拉纳(Maturana)、麦卡洛克和皮茨(1959)关于蛙眼视觉系统语义功能以及"蛙眼告知蛙脑什么"的研究,这是研究神经活动表征特征必然会涉及的问题——这一问题最初在麦卡洛克和皮茨发表于 1943 年的论文中就已提出。

4.2 神经网络与命题逻辑

我们了解到，早在1890年以前威廉姆·詹姆斯就曾提出，特定时间内大脑某一特定点的激活程度取决于输入这个特定点的所有其他点的激活量的总和。每次输入的强度与如下特征成正比：

1. 向特定点输入刺激的次数；
2. 每个输入刺激的激活强度；
3. 没有其他争夺激活量的竞争点。

条件1和条件2似乎是指各点之间的连接强度，随共时激活频率和强度的增加而加强(回想詹姆斯对于局部记忆的**频率**和**强度**联想原则)。因此，每次输入都有一个强度阈值或者说"权值"(回顾第1章"詹姆斯神经元"的相关内容)。如果特定点是所有输入量的总和，那么条件3似乎是多余的，因为被其他竞争点争夺了的激活量完全可以不计入输入。然而，我们也许可以这样认为，这种"争夺"正是一种原始的"抑制"形式——在加和过程中具有否决权。这正是我们将要探讨的另一种神经单元，即所谓的麦卡洛克-皮茨神经元。 80

麦卡洛克和皮茨

一般认为，最早有关神经系统形式计算特征的研究是麦卡洛克和皮茨(下文简称M&P)于1943年发表的《神经活动内在概念的逻辑演算》。即使是按照现在的专业标准看，这篇论文仍然非常凝练，甚至还有些晦涩。我们将介绍他们的一些基本发现，力图能够完整地复述他们最初所持有的观点。M&P是以评论那个时期在理论神经生理学内的某些"重要假设"而开始进行研究的。为了使神经计算理论更具形式化特点，他们提出(1943:22)：[1]

1. 神经元的活动是一种"全一或一无"的过程；[2]

2. 在潜加作用(latent addition)期内，总有一些固定数目的突触被激活，以便能在任何时刻激活一个神经元，[3]这些固定数目的突触与神经元先前的活动及位置无关；[4]

3. 在神经系统内唯一有效的延迟是突触延迟；[5]

4. 任何抑制性突触的活动都能绝对阻止在那个时刻神经元的兴奋；[6]

5. 网络结构不随时间而发生变化。

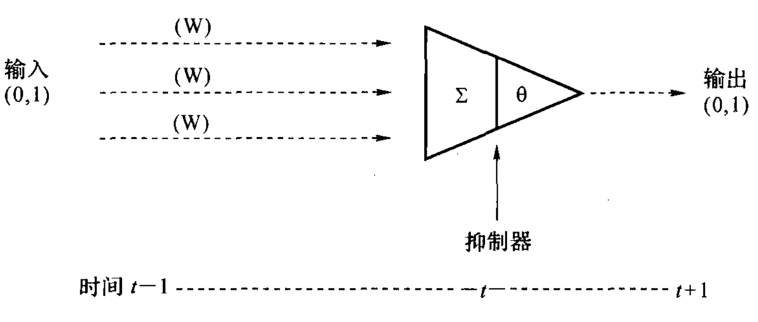


图 4.1 麦卡洛克-皮茨神经元

81 综合以上特征，M&P 神经元可由图 4.1 表示。M&P 神经元在 t 时刻内对所有输入进行加和（$\sum$）。如果此时神经元受到抑制，则不会发生任何变化。如果没有受到抑制且输入总和等于或者大于阈值（θ），则神经元被激活。神经网络与形式系统通过下面这条原理而产生联系：

（P1）

任何神经元的反应事实上都等价于神经元受到充分激活的命题（1943：21）。

也就是说，每个神经元都可以指派一个命题，充分条件是满足神经元能够被激活。

（P1′）

每个神经元都可指派一个形式命题：条件是能够满足神经元的激活。

当神经元被激活时，命题为真；不能激活，命题为假。因此，神经元“全一或一无”的特征对应一个真值命题。进而得出：

（P2）

存在于神经活动内的心理关系与……命题关系相对应（1943：21）。

因此，神经系统的“开关回路”与二值逻辑命题具有映射关系。M&P 构造了一个形式系统模型模拟神经活动，并且提出了有关这种模型的一些运算规则[7]。那时，人们对神经系统自身的结构还很不了解[8]，因此他们对神经网络的图示表征产生了较强的影响力，以这种或那种形式出现在之后的

文献中[9]。这些图示表征，如图 4.2，神经元 2 的激活表示神经元 1 已被激活[10]；图 4.3，神经元 3 的激活表示神经元 1 或者神经元 2 被激活；图 4.4，神经元 3 的激活表示神经元 1 和神经元 2 同时被激活。

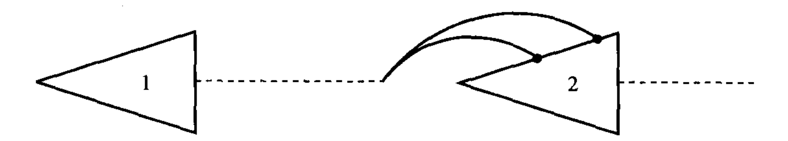


图 4.2 “时滞”网络（来源：McCulloch and Pitts, 1943, 图 1a）

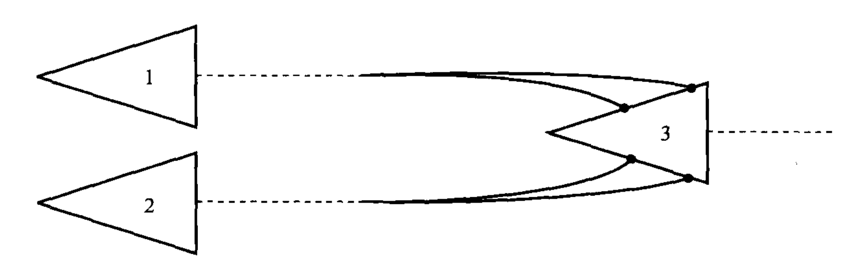


图 4.3 “或”网络（来源：McCulloch and Pitts, 1943, 图 1b）

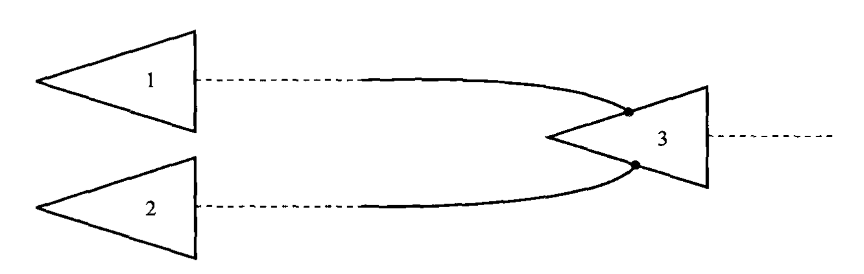


图 4.4 “与”网络（来源：McCulloch and Pitts, 1943, 图 1c）

然而，这些图示并不能明确地表达抑制关系，也没有表示出阈值。随后，M&P 神经元和网络图示得到某种程度的简化。抑制用空心点表示（如图 4.5）。图下方表格表示每一单元的“激活规律”。因为 M&P 神经元可以有很多输入，因此能够编排成任意复杂的模式。M&P 认为，当给予这种神经网络以记忆功能时，这些神经单元（包括环路）网络能够具有很强的计算能力： 82

83

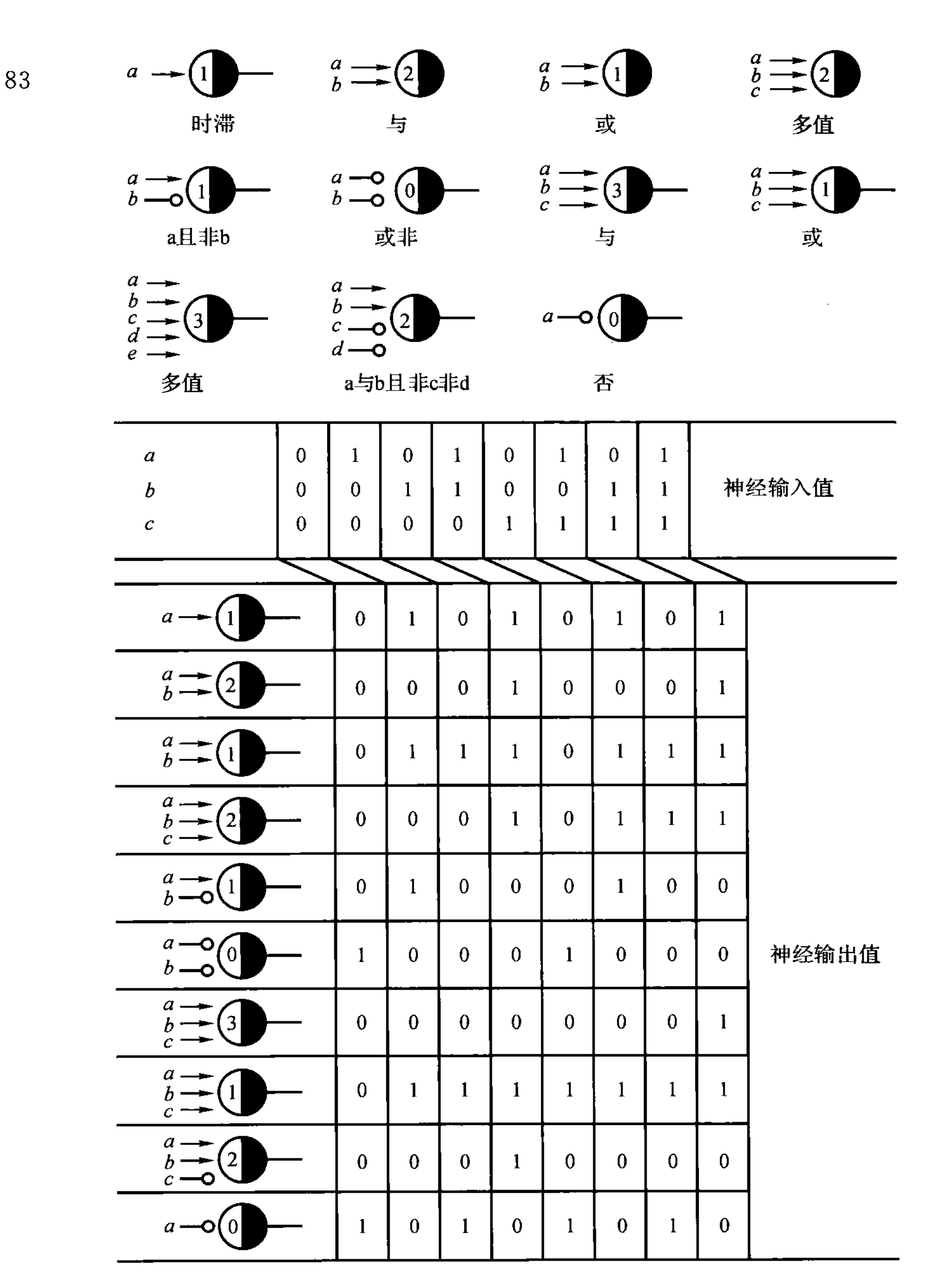


图 4.5 M&P 神经元(来源:Minsky, 1967:35,图 3.1-1)

(P3)

首先，如果为每个神经网络配备一条纸带，一些与传入神经[输入]相联系的扫描器和适合于完成必要操作运算的传出神经[输出]，那么它就只能计算如图灵机所能计算的那些数字；第二，任何后面的数字[即可由图灵机计算的]都能由这样的神经网络计算(1943:35)。 84

图灵机除去磁带(存储记忆)的部分所进行的都是有限状态控制，所以这条原理说明，M&P 网络等价于有限状态的自动机。我们将在第 6 章继续介绍图灵机和它的计算能力。

心理推论

M&P 用一些概括性(同时有些晦涩)的语言评论了这种神经网络的认识论意义。其中有一点，他们说道："就心理学而言，无论它是如何定义的，对这种神经网络的研究会为这一领域可能取得的所有成果作出贡献——即使这种分析最终指向了心理单元或'心理原子(psychons)'，因为心理原子恰恰就是单个神经元的活动。既然这种活动具有内在的命题特征，所有的心理事件也就具有了意向的或者'符号的'特征。由于这些活动的'全一或一无'规则，以及它们之间的关系与逻辑命题之间的关系所具有的一致性，所以可以肯定心理原子之间的关系就是二值逻辑命题之间的关系。**因此在内省的、行为的或生理的心理学中，基本关系就是二值逻辑关系**"(1943: 37-8)。

它被证明是一个非常有影响力的原理，然而有理由一定要接受它吗？首先，并不清楚如何能够从"全一或一无"的神经活动中，推出心理学中的基本关系就是这样的二值逻辑关系。显然，神经元**或许可以**被指派命题，如 M&P 所指派给它们的一样，它们之间的关系也**或许可以**由命题关联而形式化，但这并不是必然的。例如，我们都知道思维与神经活动模式相对应，但神经活动模式由统计定义，并非由其构成要素的布尔函数定义。其次，与单个神经元和由单个神经元构成的神经网络相联系的命题，并没有能够进行思维的适当的语义特征。人们一般所想的命题都涉及人物、地点和事件——很少考虑神经元激活需要满足的条件。所以，即使逻辑命题能够模拟神经元或神经元的激活，但却不能模拟在具体实例中思维如何激活。第三，从他们的文本中，并不清楚思想(思维)、形式网络、神经网络与被指派到这样网络的命题之间有什么关系。有时他们似乎认为只有神经活动才能实例化思维，而在另一处又有关于"网络"心理学这样的中立说法。这里有一个颇具争议的问题，即"多重可实现性"或者"多重实例化"问题——思维的 85

产生取决于网络(逻辑)**组织**,还是取决于网络构成的**材料**。如果是后者,那么思维只能在神经系统(或者因果等价系统)中产生;如果是前者,那么只要能以满足充分因果关系的方式组织起来,任何材料都可以拥有思维能力。从这一特殊意义上讲,如果一台硅制机器包含具有可实践的恰当网络,那么这台机器也就可以进行思维。对于这一点,即(思维实现)硬件在适当范围内并不是关键,这也许只是这个原理产生重要影响的一个方面。

4.3 感知器

> 感知器第一次被描述的时候就带有能够产生感知的特征。它是第一个精确的、具体的、计算导向的神经网络,同时对很多领域都产生了巨大影响。
>
> ——安德森和罗森菲尔德(Anderson and Rosenfeld, 1988: 89)

感知器的研究标志着联结主义在历史上出现了双重拐点,其一与当时"控制论"运动有关,其二与当时正在发展的数字计算运动有关。第一,感知器最初由弗兰克·罗森布拉特(Frank Rosenblatt)提出,他为含混不清的"控制论"运动增加了一些非常必要的原则。如他在早期著作中写道:"那些理论家……普遍缺乏精确的形式阐述,分析也不甚严谨,以至于常常难以看出他们所描述的系统能否在现实的神经网络中运行……与熟练网络布尔代数的分析者相比,缺乏一种可靠的分析语言已经成为一个最大的障碍。我们这个小组的贡献可以看作是提供一种应该探索和研究什么的建议"(1958: 389)。第二,我们还将了解到,感知器随后受到明斯基和帕佩尔特(Minsky and Papert,1969)的严格审查和摧毁式的批评,因此人们开始反对20世纪50年代"神经一逻辑"取向,而转向了60年代和70年代的数字计算取向。

罗森布拉特

> 信息和记忆以何种形式储存?储存库或者记忆中的信息如何影响认知和行为?……这里提出的理论,采用经验主义或者"联结主义"的立场对这些问题进行了回答。
>
> ——罗森布拉特(Rosenblatt, 1958: 386-7)

86 在麦卡洛克和皮茨发表他们的研究成果大约10年后,弗兰克·罗森布拉特和他的工作小组研究了一种称为"感知器"的装置,被看作是一种"可变

连接的麦卡洛克-皮茨网络"[11]。罗森布拉特最初提出感知器(一种"神经系统或者机器假说",1958:386)的观念是为了反驳"感觉信息以编码表征的形式储存"的"数字计算机"的观点(1958:386)。罗森布拉特认为,"这个假说所具备的易明性和智能性是非常吸引人的"(同上),更重要的是它引出了"一种内容丰富的脑模型,相当于可以执行特殊运算法则的简易逻辑人工制造物"[12](同上)。罗森布拉特继续写道:"之前的所有模型在与生物系统相匹配的一些重要方面都不成功(缺乏等势性,缺少神经经济性,依赖于过多的具体连接和激活同步性,激活单元的充分刺激不可实践,假设变量或者功能特征没有已知的神经对应物等等)"(1958:388)。正如我们将看到的,这段话读起来颇像当代的"联结主义优点"。按照罗森布拉特所言,无论对计算机模型作怎样的调整都不能解决这些难题,"它们在原则上的区别已经显示得足够清楚"(同上)。所需要的是,"网络分析员们"做一种转变,这个转变能提供一种语言,"从对系统中事件所做的数学分析来看,系统的组织结构只能够被粗略描述,它的准确结构是未知的"(1958:387-8)。之后,罗森布拉特提出了他的感知器模型假设:

> 1. 在神经系统中,涉及学习和再认的物理连接在不同的有机体中是不相同的。
>
> 2. 细胞连接的起始系统具有一定程度的可塑性。
>
> 3. 当大样本刺激呈现给神经系统时,那些"相似"的刺激(某种意义上需要精确的物理系统语言定义)会形成某种指向具有相同反应的细胞集通路;那些明显"不相似"的刺激会形成指向具有不同反应的细胞集连接。
>
> 4. 积极与/或消极的强化作用……能促进或者阻碍当前正在发展的任何连接形式。
>
> 5. 因为相似的刺激具有激活相同细胞集的趋势,所以在系统中**相似性**可以在神经系统中的某种层次上得到表征。(1958:388-9)

87

这些假设大多数都被当代联结主义系统接受,尽管在"可塑性"和"连接形成"的功能解释上,当代联结主义者使用的是连接强度发生变化,而不是硬件发生演化的术语。

感知器结构

罗森布拉特(1958)提出了两种感知器结构,一种具有三个连接层和四个单元层,另一种具有两个连接层和三个单元层,他重点阐述了较为简单的

后者。四层(单元)的感知器包括“视网膜”、A-I 投射区、A-II 联结区和一组反应集。第一层的连接是固定的,第二层和第三层的连接是随机的(见图 4.6 和图 4.7)。

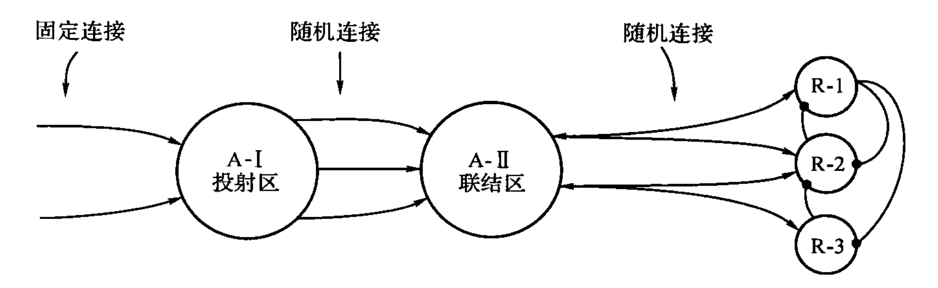


图 4.6 早期感知器的组织结构(来源:Rosenblatt, 1958:389,图 3)

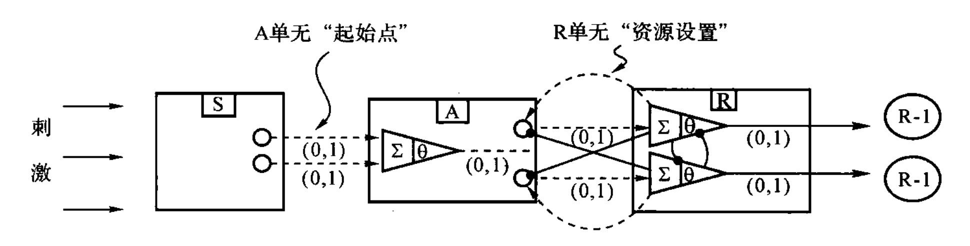


图 4.7 三层(单元)感知器的组织结构

感知器具有五种基本结构规则(这里略作简化):

1. 刺激映入视网膜,以全一或一无的方式进行反应。

88

2. 冲动传入(A-单元)细胞集 A-I 和 A-II 中(A-I 单元细胞可忽略)。

传递冲动到特殊 A-单元的视网膜点集称为 A-单元的**起始点**,这些起始点要么兴奋,要么抑制。如果兴奋或者抑制冲动的总和等于或者大于阈值,那么 A-单元以全一或一无的方式激活(见图 4.8)。[14]

3. 投射区与联结区之间的连接随机。

4. 反应单元的反应与 A-单元的反应一致。图中箭头的指向表明,到 A-II 的传递为前馈式,但 A-II 与反应单元间的传递却是反馈式。一般而言,反馈要么使其发生源(source-set)兴奋,要么使之部分地受到抑制,感知器是典型的第二种类型。这样,系统的反应便是相斥的,因为如果反应-1 发出反馈,则将抑制反应-2 和发生源,反之亦然。

5. 为了使感知器能够学习,需要在一组刺激在反应-1 发生源比在反应-2 发生源能够激发更强的冲动时,能够对 A-单元或者它们的连接进行修正。

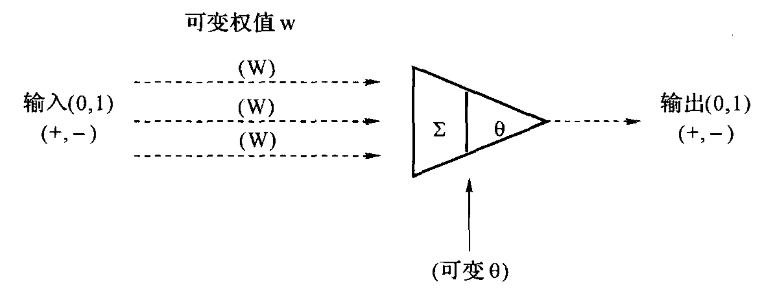


图 4.8 罗森布拉特神经元

感知器 Mark Ⅰ

如其他神经一逻辑模型一样，感知器可以通过在计算机上模拟或者实际建造进行研究。感知器 Mark I 就是一台现实机器，由康乃尔大学航空实验室建造。它的视网膜是一个 20×20 的光电管网栅，能够对呈现的不同图片进行分类。它们随机地(一个光电管最多可连接 40 个联结单元)连接到 512 个联结单元，之后再与 8 个输出 1 或者 -1 的二进制反应单元相连接。这台感知器 Mark I 能够以某种非常有趣的方式进行归纳，学会多种不同的分类方式(参见：Block，1962，包括参考文献)。

简单感知器的训练："感知器收敛过程"

想象我们正在训练一个感知器区分男性(M)和女性(F)的图片。如果我们给它一张 F 的图片，它作出了 F 的反应，则 A-单元和反应-单元间的连接权值不作变更。如果给它一张 M 的图片，它作出了 M 的反应，同样如此。但如果给它一张 F，它以 M 反应，则更改 A-单元和反应-单元间的激活连接权值(如果反应为 -1 则降低，如果 +1 则升高)。如果给它一张 M，它的反应是 F，重复上述过程(Block，1962：144)。

> *感知器收敛定理*(简化)：如果 F 和 M 的类别线性分离(见后文)，那么必然存在一种简单感知器，只要给予足够训练，它就能够学会区分这些类别(Block，1962：145，定理 9)。

评估和总结

罗森布拉特分析了很多不同感知器的行为，提出了一些我们非常感兴趣的结论：

1. 在一个**无差别**、随机的环境中，感知器能够学会将一个特定反应与一个特定刺激联结在一起，但随着学习刺激数目的增加，正确反应率减少，感知器不可能学会归纳。

90

2. 在一个**差异**环境中，每一种反应都与一类显著相似的刺激相联结，正确反应率随着联结细胞数目的增加而提高，之前没有呈现的刺激被正确分类的概率随着具有相同模式的刺激数量增加而增加。

3. 感知器的存储为**分布式**。在此意义上，系统中大部分细胞都可被每一种联结利用。除去系统的任意一部分，感知器的任何一种辨别或者联结效果并不会明显消失。但是，对比**完整状态**下的学习联结，感知器开始表现出某种整体性上的缺陷。(Rosenblatt，1962:405)

总结

我们从罗森布拉特的工作中，能够看到一些当代联结主义的重要议题：

1. 批评数字计算取向脱离生物现实；
2. 神经元的活动结构；
3. 强调统计与逻辑方法；
4. 强调再认和学习模式；
5. 运用并行加工和分布式表征。

20 世纪 80 年代早期，联结主义者发现了某些改进简单感知器局限的方法，对于(神经)网络的研究在当代也正处于繁盛时期——即使明斯基和帕佩尔特仍怀疑对于他们当初指出的感知器的局限能否进行小的修补就可以进行时下的研究(参见 Minsky and Papert, the Epilogue: The New Connectionim, to the 1988 edition of their 1969 book)。我们将在第 11 章介绍联结主义计算机器时回到这个问题的讨论。

4.4 线性分离与 XOR:麦卡洛克-皮茨网络与感知器

XOR 连接和它的否定形式表明，以类神经单元网络组织的(简单)感知

器，在计算能力方面存在着一个简单而又重要的难题。

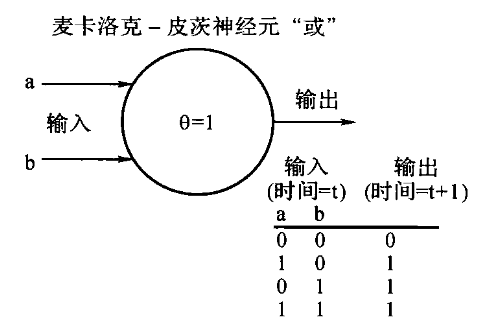


图 4.9 简单 M&P“或”神经元

（来源：Anderson，1995：49，图 2.9；经 MIT Press 许可复制）

线性分离

91

想象一个单元有两个输入 P 和 Q，以及一个阈值。如果输入的总和大于阈值，输出 1，否则输出 0（如图 4.9 所示）。假定输入是其他单元的输出，那么输入也要么是 1，要么是 0。我们假定这个单元的阈值是这样的，如果 P 和 Q 都是 0，则输出 0，其他情况则都输出 1（这是 OR 规则）：

P	OR	Q
1	1	1
0	1	1
1	1	1
0	0	0

现在使每个输入线为一个“维度”，两个输入线也就是两个维度。OR-单元的所有可能输入可以由一个二维平面表示。我们把输入条件的开（1）或关（0），用 x、y 的真值函数标出，其中 x 轴是第一组输入值，y 轴为第二组输入值，所以 OR 可由图 4.10 表示。在这个图中，我们可以画出一条直线，将平面分割成关（0）和开（1）两部分（图 4.11）。这说明，OR 是线性分离的。16 个基本的真值函数中有 14 个是线性分离的——只有 XOR 及其否定形式（见后文）不是。

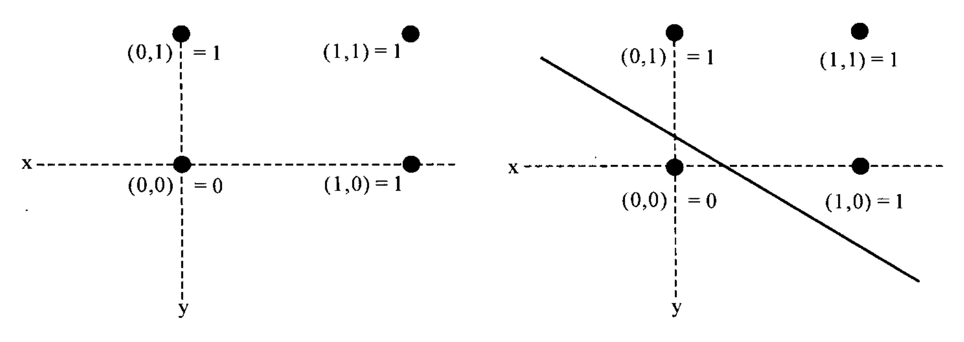


图 4.10 “或”的笛卡尔坐标　　　　图 4.11 “或”的线性分离

XOR

二值函数中异或(XOR)结构和它的否定形式不是线性分离(如图
92 4.12)。XOR 和它的否定形式的非线性分离,见图 4.13。XOR 需要两条线才能划分开(1)和关(0)状态,这就表明 XOR 是非线性分离的。XOR 的否定形式只需把图中 0 和 1 状态交换位置,显然也是非线性分离。

P	Q	P 异或 Q	非(P 异或 Q)
1	1	0	1
1	0	1	0
0	1	1	0
0	0	0	1

图 4.12 “异或”及其逻辑“非”真值表

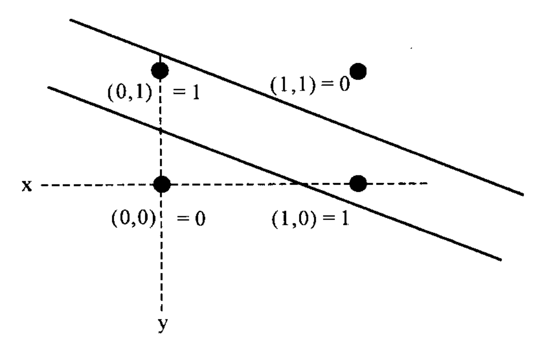


图 4.13 “异或”的非线性分离

线性分离的概念同样适用于多于两个输入单元(多于两个维度)的情
93 形。例如,一个单元有三个输入线:P、Q 和 R。如果对之作图,则需要一个

三维空间,像一个立方体。但是一条线不可能把一个立方体分成开和关两个区域,这就需要一个平面才能做到。人类的空间直觉不能在空间上想象多于三个输入线的情形,但在形式上都是一样的,能够划分那样一个高维空间的面称为"超平面"。更精确地说,"超平面"的等式是:**在这个空间中,突触权值与单个输入乘积的加和等于阈值的点集合**。如果输入空间存在一个超平面,并且输入类别是线性分离的,那么原则上是可被学习的。

麦卡洛克-皮茨网络和 XOR

麦卡洛克-皮茨单元能够计算两个输入的真值函数 16 个命题中的 14 个,两种不能计算的是异或命题(XOR)及其否定形式。然而,只要给两个输入的单元连接一个 AND 单元,就可以计算 XOR(及其否定形式)了。当且仅当两个输入都为开(on)时,AND 单元使系统关闭(off)。

感知器和 XOR

前面提到,如罗森布拉特认为,感知器是带有一层可变连接的一种特殊 M&P 网络结构,且这种装置满足感知收敛定理(PCT)。既然 M&P 网络可以计算 XOR,那么感知器能不能呢?确实有一些"感知器"网络可以计算,如图 4.14(其中 Ψ 的阈值是 1)所示的例子。在这个网络中使用⟨1,1,−2⟩的连接权值,它可以通过训练过程使感知器学习吗?如果可以,那么这个例子是否可以说 XOR 是感知器通过训练过程不能学习非线性分离函数的反例呢?(如果不能,是否可以说 XOR 是感知器可以计算的函数,但却不能知道如何学会计算的例子?) 94

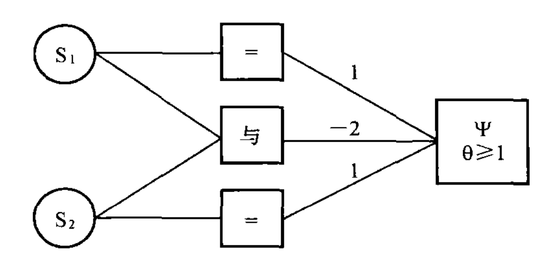


图 4.14 "感知器"的"异或"计算(来源:Quinlan, 1991:33,图 1.11)

结果是肯定的,这个感知器能够学习⟨1,1,−2⟩,这就是一个通过感知训练过程能够学习的非线性分离函数的例子。使第一层连接权值保持不

变,**默认所有的连接权值一致都为1**,即⟨1,1,1,1⟩。这些值不能通过感知器训练过程学习,当与可变连接层的权值组合时,它们不能保证通过训练过程能够学习。按照感知器收敛定理,如果一个函数是线性分离的,那么通过足够的训练,感知器必然能够学习它。我们现在看到了一个相反的情况:如果一组数据是**非线性分离**的,即使经过足够的训练过程,感知器**并不必然**最终能够学习它们。感知器可能会碰巧偶然地找到解决办法,如上面XOR的例子,但是并不必然如此。

而且,这个感知器尽管也许在连接不变层使用一组⟨1,1,1,1⟩的权值能够学习XOR,但是当第一层的权值要求**不是**⟨1,1,1,1⟩时,它就不能学习任何函数命题了。这是因为,当一个函数要求第一层出现其他的值,而非一致的都是1的时候,就不会被这个感知器表征,也不会通过训练过程学会。如果人为地参与改变权值,使得感知器在可变层能够学习正确的权值,这样也必然会出现一些感知器不能学习的其他函数(如XOR)。

4.5 简单探测器语义

95 所谓"简单探测器语义(simple detector semantics, SDS)"是指,一个神经过程(或者更普遍的任何"神经网络")关涉什么启动了这个神经过程——神经的激活显示了或者**探测到了**神经元将要调整到的特定状态:当指定的特征(目标、特性或者环境中事物的状态)出现时,单元或者单元集就会**开启**[15];当出现其他特征时则**关闭**。可概括为:

(SDS)

单元或者单元集(神经元或者神经元集)表征能够开启它们的事物。

这个概念可由多个名称表达,包括"指示器"、"协变"、表征的"因果构成"以及"信息"理论。在某种程度上,SDS的提出受到了神经元单根纤维和单个细胞研究的启发。最初使人们深入理解这个观点的是勒特文等人的工作,以及他们的经典论文——《蛙眼告知蛙脑什么》。

勒特文、马图拉纳、麦卡洛克和皮茨与《蛙眼告知蛙脑什么》

为了理解"蛙眼告知蛙脑什么",有必要了解有关蛙的一些特征。

蛙

1. 蛙依靠视觉捕食。

2. 它的眼睛和人类的眼睛一样，看到猎物时眼球紧跟不动，而身体动。因此它的眼睛非常稳定，在其视觉范围内没有中央窝或者特别敏锐的区域。

3. 蛙只有一个单一的从视网膜到脑丘的视觉系统，与人的双重系统不同。

4. 它不明白，也并不在乎它周围世界的静止部分是什么样的。

5. 如果昆虫不活动的话，蛙也会在周围满是食物的情形下饿死。

6. 它会跳起来去抓类似蠕虫和昆虫大小的物体，只要这种物体看起来像蠕虫或者具有昆虫那样的活动，因此蛙很容易受欺骗。

7. 它通过跳到较黑暗的地方来逃脱危险。(Lettvin et al. ,1959:231)

探测器

蛙从眼睛到大脑的路径，包含四个并列的连接到脑接收区域的组织构造：　96

1. ***对比探测器***　能够察觉到在最小视觉区域中，具有明显轮廓的事物是运动的还是静止的，或多或少通过对比来实现。

2. ***凸面探测器***　察觉在稍微大一点的范围内，如果目标颜色比背景要多少黑一些且正在它上面运动，那么目标是否具有曲线轮廓；只要轮廓属于那块区域，并且较为明显，当目标停止时它将记住这个轮廓；如果这个目标间歇性地朝着某一背景移动，那么就会加强对这个目标的关注；如果有一个阴影遮住这一目标一段时间，那么对这个物体的记忆将逐渐消失。

3. ***运动探测器***　察觉在其视域且较大的区域内是否有一个移动的轮廓。

4. ***模糊探测器***　察觉在最大区域内，还存在着多少看起来模糊暗淡的目标，主要通过与目标的距离和它们的运动速度进行权衡。

"昆虫探测器"

凸面探测器(上面所述构造 2)证明具有一些非常有用的特征："当黑色物体进入视域并做间歇性运动时，这种组织构造(构造 2)会发生明显反应。即使光线发生变化或者背景(比如说花或者草的图片)移动，这种反应都不会受到影响。还能找到一个比这个更好的**检测昆虫**的系统吗?"(1959:253-254)勒特文等人这样描述探测器语义："在蛙的视神经中，这种特殊的组织

构造用来测量什么呢？**我们认为，它是用来测量具有某种特征的刺激数量的。这种刺激能够使这种组织构造达到最大限度的兴奋，可将组织构造的兴奋称之为特征**”(1959:253)。在该文的结尾部分，作者用检测器的激活命名特殊的刺激特征；也就是说，**检测器的激活表征出现了那种刺激特征**。该文前面部分则说明，至少在蛙的生存环境中，通过检测这种刺激特征，蛙可以知道什么物体可被认为是昆虫。人们发现，在猫的视觉皮层中也同样存在一些神经元，当接受水平光柱时会被激活，而另外一些神经元只有在接受垂直或者特殊角度的光柱时才会被激活。我们也许可称之为定向“边缘检测器”[16]。后来，许多文章又报道发现了许多其他类型的检测器[17]，然而同时也出现了一些相反的证据[18]。

97

简单探测器语义难题

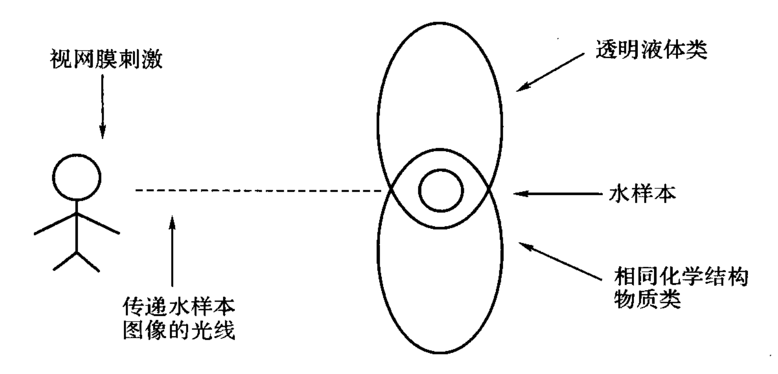
“构造组织 2”的激活意味着**探测到了昆虫的信号，昆虫开启了“构造 2”的开关**。任何理论在其起始阶段都注定会遇到很多难题，这种理论也不例外。这里，我们将重点讨论初级探测器语义遇到的如下难题。

真正因果或者纵向难题

太阳发出的光子，会被大气中的颗粒反射，被草或者树木反射，被昆虫反射。按照 SDS 理论，在“构造 2”激活的过程中只有被昆虫反射的光子才会被蛙的视网膜吸收，转换为电冲动，又变成其他电冲动引起“构造 2”的激活。问题：这里还有很多其他的因果因素，为什么只说是昆虫使蛙的“构造 2”被检测激活？为什么不会是视网膜，或者树木、阳光的原因呢？是什么决定了正确的**纵向**因果链条？

横向或者质性难题

如果我们想知道“构造”因为昆虫而激活的相关因素，那么如何验证哪种有关昆虫的特征是有效的呢？毕竟，就通过反射光线而言，完全可以有很多其他的昆虫特征，如：**有翅的东西，害虫的潜在的恐慌，带有杀虫剂的潜在药性，来自 9300 万公里之外不规则运动的光点**等等，但为什么只认为是由于寻找到了这种反射光线的特征（认为是找到了昆虫）而被激活，而不是因为其他的可供选择的特征呢？图 4.15 说明这两个问题。



纵向难题：在某一位置上的某物看起来有一定的"深度"，而当我们想到它是在"水"中时，就能解释这一现象。水只是透明液体的一种，为什么我们会选择"水"呢？横向难题：水有多种属性，为什么我们只选择了它的透明属性呢？

图 4.15　纵向和横向难题

（来源：Braddon-Mitchell and Jackson，1996：69，图 3；经出版社许可复制）

错误表征或者析取难题

第三个是错误表征难题：错误表征为什么是可能的呢？（这个问题最初由福多提出，标以"析取问题"而闻名[19]。）根据 SDS，蛙是这样的：当出现一只（移动的）昆虫时，它的"组织构造 2"被激活，然而激活它的也可能是一块苍蝇大小的磁铁："把一张大幅有花有草的彩色照片放在蛙的视觉范围内，在蛙看来完全是一片自然生活的场景。在蛙'组织构造'的接收范围内——7 寸距离，移动图片，蛙没有任何反应。取一块苍蝇大小的磁铁放在图片上，在蛙视域内移动，我们发现蛙表现出了明显的反应。但如果把磁铁固定在图片上，与图片整体移动，蛙却没有反应"（1959：242-3）。据此，我们需要注意这里面究竟发生了什么。最初我们从蛙的例子得到两点认识：首先，简易探测器语义理论（SDS），即构造 2 的激活表征能够开启它的事物；其次，"构造"是一种"昆虫检测器"。但是，我们接着发现"构造 2"同样对不规则运动的磁铁（moving magnet，MM）反应，这就是存在的问题。"构造 2"被称为"昆虫检测器"，但是它却也对 MM 反应，那么根据 SDS，它**也**可称之为"MM 检测器"。那么，它对昆虫或者对磁铁都有反应，或可称之为"昆虫－或－磁铁检测器"。然而，如果它（真的）是一个"昆虫－或－磁铁检测器"，那么它会正确地把 MM 认作是"昆虫－或者－磁铁"，不会把 MM 错误地表征成昆虫。这就**避免了错误表征的可能**了。然而，错误表征还是可能的，1－或－2

98

是错误的。简言之，一个潜在的错误表征可以被这个理论(SDS)转变成一个析取式的正确表征，因为出现昆虫或者磁铁的时候，这个“构造”都会被激活。事实上，就此事而言并没有理由认为称“构造 2”为“昆虫检测器”(但它会误测磁铁)不如称“磁铁检测器”更好(但它会误测昆虫)。

99 我们怎样反驳呢？(1)可以认为勒特文等人错误地把语义构造 2 说成了“昆虫检测器”——而不是**不规则移动的小黑点探测器**——这便是第一次引进“凸面检测器”的原因(但也并非准确)。如此的话，MM 例子完全不是因为错误表征，这里并不存在疑难。(2)我们可以修改 SDS，加上有关蛙生活的有关“自然环境(natural environment)”：

(SDS-NE)

单元或者单元集(神经元或者神经元集)表征在其生活的环境系统中能够开启它们的事物表征。

这种修改了的构造 2 就可以保证蛙能够在充满昆虫(而不是 MM)的环境中准确地对食物进行定位，而得以生存。无论哪种改动，都不得不对原来的观点作出修改。

注释

[1] 这个“重要假设”没有反映出现代神经科学中“信号速率与轴突直径成正比”的事实。

[2] M&P 主要讨论动作电位的“全一或一无”特征，已被证明(归因于后来更加娴熟的技术)以小于毫秒的速率递级激活传递。“重要假设”中也没有涉及神经递质在激活传递中的作用。

[3] 事实上，每个神经元都有一联结阈值，必须满足或大于这个阈值才能激活神经元。

[4] 事实上，所有的输入都有相同的联结强度或权值。

[5] 事实上，这是一种不应期，在这期间神经元不能够被激活。

[6] 注意到单元对兴奋性和抑制性输入的加和，与随后的感知器和大多数联结主义网络中的神经元相同。

[7] 这个系统建立在怀特海和罗素(Whitehead and Russell，1925)，以及卡尔纳普(Carnap，1937)的研究基础上。

[8] 例如，明斯基(1967：36)说：“麦卡洛克和皮茨的论文最早受到赞赏的并不是因为它的标记方法，而是因为它有关哲学和技术上的内容”。

[9] 例如，麦卡洛克-皮茨的图示被冯·诺依曼在 1945 年的 EDVAC 报告中采用——随后使用的人也越来越多。

[10] 在这些图示中，两个实心点表示神经元的兴奋，对比“与”和“或”的图示。换句话说，这暗示了阈值是 2。

[11] 参考 Cowan and Sharp(1988)。

[12] 在这里，罗森布拉特提到了明斯基及其他一些人的工作。也许是在阅读了罗森布拉特对“逻辑人工制造物”的研究后，激发了明斯基随后的评论。

[13] 我们在这里计算的是实际联结层，而不仅仅只是可调节层。

[14] 罗森布拉特神经元首先对它的权值输入进行加和，如果加和结果等于或大于阈值，则输出；如果结果不等于或小于阈值，则静息。 100

[15] 或者它是得到更多的激活。神经元开启或者获得更多的激活，都用“开”表示；相反地，则用“关”表示。

[16] 参考 Hubel and Wiesel(1979)。

[17] 参考 Gross et al. (1972)的短尾猿“手臂探测器”。

[18] 参考 Stone(1972)。

[19] 参考 Fodor(1984,1987)。Cummins(1989)的第 5 章命名为“错误表征难题”，我们引用了康明斯的这种表达。

【思考题】

麦卡洛克和皮茨(M&P)

M&P 对神经系统的形式化提出了哪五个假设？

这五个假设相对于 20 世纪中期有关神经元的研究有多精确？

神经激活的哪种“特征”可以适用于命题逻辑？

形式系统的连接神经网络的两个基本原则是什么？

为迟滞、与、或绘制 M&P 单元。

当提供存储时，M&P 网络的计算力是指什么？

M&P 论文具有怎样的历史意义？

M&P 论文关于心理推论的三个潜在问题是什么？

罗森布拉特：感知器

罗森布拉特反对“计算模型”的五个指责是什么？

关于神经系统建构感知器的五个假设是什么？

最初的感知器

最初的感知器具有怎样的组织结构：什么是单元层，什么是单元层间的各连接？

基本(简单)感知器

基本的(简单的)感知器具有怎样的组织结构：什么是单元层，什么是单元层间的各连接？ 101

可变联结层有多少层？
在感知器中，反应单元之间相互抑制吗？
哪些层次之间仅仅是前馈的？
哪些层次之间是反馈的？
在什么意义上说，感知器的存储是“分布式”的？
在什么意义上说，感知器是“并行”加工？
一般来讲，怎样训练感知器？
罗森布拉特从他的实验中得出的三个结论是什么？
在罗森布拉特的讨论中，能找到与当代联结主义相同的哪五个议题？

线性分离与 XOR：麦卡洛克-皮茨网络与感知器

什么是线性分离函数？
说明 XOR 并不是线性分离函数。
（简单）感知器能够计算 XOR 吗？
（简单）感知器能够学习 XOR 吗？
（简单）感知器仅仅能够学习线性分离函数吗？
感知器收敛定理指的是什么？
（简单）感知器能够计算以及学习计算一种感知器收敛定理并不保证它能够学习的函数吗？

简单探测器语义（蛙眼告知蛙脑什么）

什么是“简单探测器语义”？
102 在蛙实验中，发现了哪四种探测器？说说它们分别做什么？
在蛙实验中，对发现的四种探测器来说，接受域的结构是怎样的？
哪种探测器被称为“昆虫探测器”，为什么？
正如勒特文等人所表明的，如果蛙也尝试去吃移动的磁铁，为什么还称之为“干扰器”？
“错误表征（或者析取）难题”是指什么？
这个问题是怎样被提出的？

【推荐读物】

概论

Anderson and Rosenfeld(1988)是同时期关于神经逻辑问题最好的论文集，包括了大量的有关问题和有用的介绍文章。在 Quinlan(1991)的第 1 章，可以找到关于麦卡洛克-皮茨网络、感知器和 XOR 的有价值的总体看法及讨论，这一章还附加了对联想与联结主义之间关系的讨论。另一个较好的简短讨论出现在

Cowan and Sharp(1988)。有关这段时期,最新的简明讨论可以在 McLeod et al.(1998)中找到。Anderson and Rosenfeld(1988)中包含一些关于神经模型的先驱们令人感兴趣的总览。

麦卡洛克和皮茨

Minsky(1967)的第 3 章,对麦卡洛克-皮茨网络进行了详细讨论,前面所列的大多数文献同时也包括了对 M&P 网络的讨论。

感知器

除了上面所列的研究以外,在文献中还有大量的简短的对感知器的讨论。例如,Wasserman(1989)第 2 章,以及 Caudill and Butler(1993)第 3 章。不幸的是,他们因对感知器的一些特征持有异议而受到指责。回顾并考察 Rosenblatt(1962)中的讨论是非常有用的。Block(1962)第 1—8 章对感知器作为神经刺激的计算设置作了清晰介绍。Nilsson(1965/90)对感知器的学习作了全面讨论。



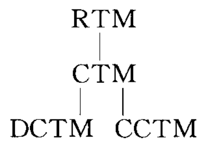
第二部分

心智的数字计算理论



引 论

第二部分的核心主题是**心智的数字计算理论**(digital computational theory of mind,DCTM)。我们把这种理论视为更具普遍意义的**心智计算理论**(computational theory of mind,CTM)的两种理论之一(另外一种理论是**联结主义心智计算理论**(connectionist computational theory of mind,CCTM),将在本书第三部分讨论)。心智计算理论也可恰当地认为是古老的**心智表征理论**(representation theory of mind,RTM)的一种特殊理论: 105



我们将采取准历史的进路介绍这一问题。有关 RTM 的观点至少可以追溯到英国经验主义者洛克和休谟(见第 1 章),但 CTM 的理论是在 20 世纪中期现代数字计算机发明后才出现的,虽然通用程序"计算器"可追溯到 19 世纪中期的巴贝奇(Babbage)。编译程序的数字计算机能够思维、表现出智能的观点最初由图灵提出,虽然我们不能肯定他也认同这样的观点,即认为人类的思维或者智能就是数字计算。如果认为人类的认知是计算的一种特殊形式,那么也就同时坚持了这样的观点:(1)心灵在内部操作心理表征;(2)数字计算机是算法符号操作器。

第二部分重点阐述不涉及认知模型的关于数字计算的一些基础内容(主要与认知科学相关)。如果计算机在算法上操作符号,那么我们就要着 106 重讨论计算机的**操作**和**符号**这两个方面,这便是我们所要做的事情。首先,介绍具有历史意义的 SHRDLU 演示项目的数字模型(第 5 章)。其次,我们将审视一些受不同观念影响的数字结构,如图灵机、冯·诺依曼机、产生式

系统、鬼蜮模型(第6章)。我们将数字结构划分为存储和控制两个维度。再次,讨论三种流行的符号图式:逻辑(谓词演算)、语义网络和框架/脚本的优缺点(第7章)。最后,我们将系统地总结(第8章)和评论(第9章)这种由计算理论产生的心智观念。

术语注解 虽然我们使用“数字”这个词语命名这种类型的心智理论,但没有企图确切地对其进行定义(很多这样的尝试均告失败)。一些人喜欢把这种类型的心智理论称作“符号的”、“经典的”或者“基于规则的”,但这些名称都不能完全地给予概括。最好的办法是先不为之贴上固定标签,只研究那些已被广泛认可属于这种类型的心智理论的机器具有哪些重要特征。基于这种考虑,“数字”至少有这样的优点,即各种类型的机器都是受到1960—1980年期间通常称为“数字”计算机模型的影响而产生的。

5 人工智能模型范例:SHRDLU

5.1 引 言

早期自然语言的处理程序,如 ELIZA(通过键盘和控制器能对病人实施心理治疗)就具有智能的表现。它的程序策略是,运用关键词搜索原先录制好的应答,然后根据"病人"的状态再向病人提出问题。这种处理程序在偶然的情况下也是具有说服力的。例如,下面是一段一位计算机公司的雇员(VP)与 ELIZA 电话对话的录音记录,VP 并不知道 ELIZA 的真实身份(Boden,1977:96): 107

VP:今天早上,我可以使用计算机吗?
ELIZA:你为什么要这么问?
VP:我或许可以额外地增加一些 telcomp 的销售量。
ELIZA:你为什么这么肯定呢?
VP:我预期顾客会对这种系统的演示表示满意。
ELIZA:这对你来说很重要吗?
VP:当然。
ELIZA:在你看来这件事情很有意义了?
VP:你是在嘲讽我吗?
ELIZA:你为什么认为我嘲讽你了?
VP:没有什么理由不允许我现在使用计算机吧?

ELIZA：就请告诉我更多一些关于你预期的顾客的信息。

VP：打我电话吧，我的电话是：491-1850。

当 VP 给一个正在睡觉的同事打电话的时候，他得到了下面的对话：

VP：你为何如此戏弄我？

同事：你为什么认为我戏弄你了？

但有谁会认为 ELIZA 的谈话与人类相似呢？而且，虽然 ELIZA 的表现很引人注意（一个秘书，当他知道 ELIZA 是电脑程序时，即使当事人正与 108 ELIZA 进行谈话，也会打断让其离开），但是它的谈话却没有任何实质性内容：没有动机、没有计划、没有语义（意义）、没有交流策略、没有环境意识、没有知识等等。

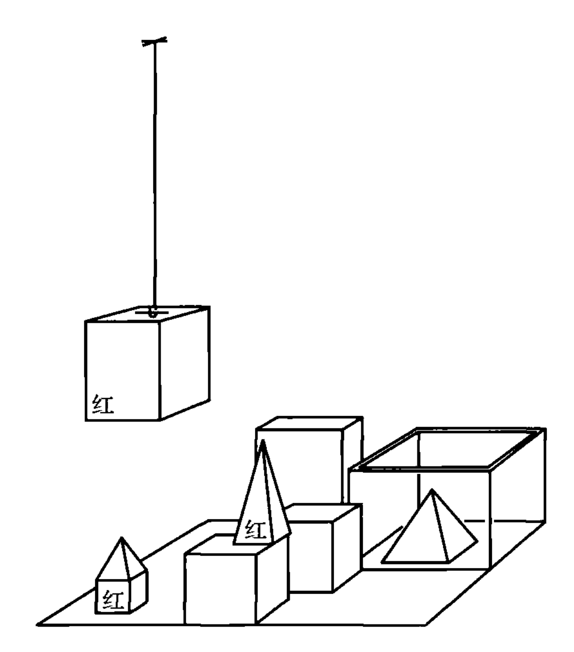


图 5.1 "拿起大的红色积木"（来源：Winograd，1972：8，图 2；经 Academic Press 许可复制）

1971 年，特里·威诺格拉德（Terry Winograd）在其 MIT 博士论文中彻底地改变了这种情况。他认为 SHRDLU[1] 经适当设计能够在有限范围内"理解（understand）"（直接引用威诺格拉德如此令人吃惊的描述）语言：在一个涂有各种颜色的积木微小世界里，有几个立方体、几个角锥和一个盒子——所有这些都放在一张桌子上。SHRDLU 能够对提出的问题、陈述和

命令进行句法和语义分析;能够对这个积木世界进行推理,并有计划地移动、应答,与对话策略和动机相一致。

5.2 SHRDLU 的对话

从下面威诺格拉德与 SHRDLU 的对话[2]中可发现,SHRDLU 似乎是非常有效的。积木世界从如图 5.1 中的设置开始。对话中的仿宋体部分是威诺格拉德输入到计算机中的文字,斜体文字为计算机的应答。对话间其他文字是对计算机展现的一些特征所作的评论。每次作出应答约需 10 秒。

1. 拿起一个红色的方块。

109

好的。

当执行这个命令的时候,系统回答"好的"。为了拿起这个红色的方块,必须将绿色方块移开,并找到一块空地放置这个绿色方块。

2. 拿起一个角锥。

我不知道你想要哪个角锥。

共有三个角锥,因此这个程序没有作出任何反应和动作。

3. 找到比你现在正拿着的方块还要高一点的那个方块,然后将它放进盒子内。

这时系统要运用推理弄清楚"那个"和"它"的具体含义,决定哪个方块要被拿起。

我假设你所说的"它"指的是比我现在所拿着的方块高的那个。

系统如上打印出来。

好的。

它的手放下红色方块,拿起一个比它高的蓝色方块。

4. 盒子里面装了什么?

蓝色角锥和蓝色方块。

5. 这个角锥是被什么支撑着?

一开始系统应答"这个角锥"时说:"不知道你说的是哪个角锥",但是现

在它假设我们所指的这个角锥就是刚刚提到的那一个。

是那个盒子。

6. 多少方块不在这个盒子里?

否定词像“没有”和“不”都会被操作,无论它们出现在句子中还是出现在词组中(如没有盒子(no boxes))。

110 *有四个。*

7. 它们之中是否至少有一个比我要求你拿起的那个小呢?

这个句子中“它们之中是否至少有一个”必须根据前面提到过的问题以及“我要求你拿起的那个”来分析,回到方块上来。有的名词词组会比较复杂,诸如“至少一个”、“多于三个”、“正好是七个”等等。先前的记忆用来推断出“哪个方块是你叫我去拿起来的”。注意到在这个复杂的句法结构中的“那个”,需要对这个句子进行两步分析才能获得。

是的,有,是那个红色的立方体。

8. 它有支撑吗? (is it supported?)

被动语句不使用“by...”的句子表示。

是的,它由桌子支撑。

9. 这个桌子可以拿起这些方块吗?

不能。

根据语义规则这是一个没有语义的问题,必然作出否定性的回答。

10. 一个角锥能够放在一个方块上面吗?

是的,可以。

这个推理系统找到了一个实际的例子,所以它知道这个是可能的。

11. 一个角锥可以放在另一个角锥上吗?

我不清楚。

这个问题的答案是可以,但是在这个情景中没有提供这样的例子,所以对于机器来说作出回答并不是一件容易的事。

12. 将两个角锥重叠起来。

我不能。

机器人试了试,但是失败了。

13. 那个蓝色角锥是我的。

我明白。

这个数据加进了数据库。 111

14. 我有(own)了不是红色的方块,但我没有可以支撑角锥的东西。

我知道了。

创造了两个新的法则,说明"有(owning)"的含义。

15. 我有(own)盒子吗?

没有。

盒子能够支撑所有装在里面的东西,包括角锥。

16. 我能有(own)盒子里的什么东西?

蓝色方块和蓝色角锥。

17. 能不能将两个红色方块或者与一个绿色立方体或者与一个角锥一起重叠起来呢?

可以。

逻辑连接词像"与"、"或者……或者"受语法和语义两方面控制。一个命令只有当它看起来像一个问题时才被执行。命令的执行过程分为下面的一系列步骤:它将绿色立方体放在了红色方块上(注意到,它选的绿色立方体上面没有任何东西),然后将小的锥体从小的红色立方体上移开,然后用这个立方体按照规定执行……

18. 哪一个立方体立在桌子上?

用来支撑红色角锥的那个(the... one which)大一点的绿色立方体。

一般来说,系统满足"大一点的绿色立方体"的回答。问及"哪一个立方体",它需要作出一种更加确切的回答。机器回答使用了指代名词("the... one")和代名词(which),使对话过程更符合自然语言习惯。

19. 有没有一个大的方块在角锥后面?

有三个:大的红色方块、大的绿色立方体和蓝色立方体。

20. 在支撑角锥的绿色立方体上面放一个小的。

"一个小的"必须与一个先前的句子"一个大的方块"进行对比,因此这 112

意味着“一个小的方块”。

好的。

21. 把 littlest 的那个角锥放到它的上面。

“(littlest)”字典里并没有这个单词,但是它的意义可以从词根“little”变化得到,即“最小的那个角锥”。

好的。

22. 有多少东西在绿色立方体的上面(on)?

我不是很确定你说的“在上面”是什么意思。

你是说①最直接地摆在上面的东西?还是②任何一个处于其上面的东西?

当有很多解释且没有一个是更好的时候,系统将会向使用者询问更多信息。这种模糊不清的词组源于字典中对“上面(on)”的定义,当启发式方法不足以解决模糊概念时,对更多信息的要求就会自动提出。当代词,如“它”或者“它们”不清楚时也会发出这样的疑问。

{22a 我意思是}②。

有三个。

23. 当你把那个绿色角锥放在小的立方体上的时候,有没有接触过(had touched)一个角锥?

有的,接触过那个绿色角锥。

系统从它的记忆中作出推断,此刻涉及哪些东西以及与它相关的其他东西。这种分析包括复杂的动词时态,并且能够清楚地解释不规则动词的含义,如“have”。

24. 你什么时候将它拿起过?

当我将一个红色立方体、一个大的红色方块和一个大的绿色方块叠起来的时候。

时间是通过其时所做的一些重要工作来描述。

113

25. 为什么?

系统通过回忆前面的句子,以理解这个问题和随后的问题。

能够获得红色方块。

通过对其原始子目标树状图的回溯进行选择,系统开始对自己的动机

有所了解。

26. 你为什么要那么做(do that)?

清除红色方块上面的东西。

“那么(that)”指代问题中的行动。

27. 你为什么要清除那个(that)方块上面的东西呢?

“那个(that)”指代刚刚提到的方块。

将它放置到一个大的绿色立方体上面。

28. 你为什么要那么做?

将一个红色立方体、一个大的红色方块和一个大的绿色方块重叠起来。

29. 你为什么要那样做?

因为你让我这么做的。

它记得之前做过什么,但并不记得被要求做的命令。

30. 你是怎么完成它(do it)的呢?

“它(it)”指上一个问题中提到的行为,而不是它对问题的回答。

通过将一个大的红色方块放到桌子上;然后把一个大的绿色立方体放在上面;然后将一个红色立方体放在这个立方体的上面。

这是完成最终目标需要进行的各个子目标的描述。

31. 在你做这件事情时(while),你接触到了多少个物体?

表达时间跨度的时间指称,可用“在……时(while)”。 114

四个。

32. 那个红色立方体在你清理它之前支撑着什么?

绿色角锥。

它记得移动过什么东西,所以可以推断这个场景之前的画面。它可以理解如“在你清理它之前”这样的句子。

33. 之后(then),盒子的左边有五个方块。

“之后(then)”指上一个句子中提到的时间。

不是,只有四个:一个红色立方体,两个大的绿色立方体,一个大的

红色方块。

在前面的对话中，有关“有(owning)”的新信息作为初步值被接受。但系统知道关于这些物体特征的全部基本信息，包括方位。它能够检验一切关于位置或者其他有关物体特征的信息，之后表达对陈述的认同与否。

34. 放置蓝色角锥盒子里的方块上面(put the blue pryamid on the block in the box)。

这个句子有些含糊，但当这么理解时就不会含糊不清了。按照当下的情景，如果将这句话理解为“放置盒子里方块上的蓝色角锥(put the blue pryamid which on the block in the box)”，是没有意义的。因此，意指“放置蓝色角锥到盒子里的方块上面(put the blue pryamid on the block which is in the box)”。

好的。

35. 这里有没有一个东西(anything)比每一个(every)角锥都要大，但是并不与支撑它(it)的东西一样宽？

系统能够理解数词，如“每一个(every)”、“任何一个(any)”、“一些(some)”，也理解如“比……大”、“与……一样宽”这样的句子结构。“它(it)”指代同一个句子中的一个名词，但“它”与其所指代的名词却有不同的含义。

有的，那个蓝色方块。

36. 有没有一个尖塔(steeple){…}

对不起，我不知道“steeple”是什么意思。

115 系统很快对它不认识的词语作出反应。

37. “尖塔(steeple)”是指将两个绿色立方体和一个角锥叠起来的东西。

这样的说明标志着我们正在定义一个新的词语。

我理解了。

38. 那么现在有没有尖塔？

没有。

这个新定义用来作为识别尖塔描述。

39. 建造一个(build one)。

好的。

上面所作的定义的描述足够让系统叠出一个那样的尖塔并且能够进行识别。其中“一个(one)”指的是“一个尖塔”,然后建造计划要开始了。首先,它通过移开一个红色立方体和两个锥形体,从而得到一个大的绿色立方体。然后将两个绿色立方体和一个大的红色角锥叠加起来,满足那个描述。

40. 称那个最大的方块为“超级方块”。

好的。

可以给物体命名。

41. 从开始到现在,你有没有拿起过超级方块?

是的,有过。

42. 你为什么将它扔了?

我将你说的“它”,理解为那个超级方块。我不能够解释一个不存在的事情。

在记忆中没有记录任何一个方块曾被扔过。

43. 红色角锥的右边有没有别的什么东西?

有的,有四个:一个盒子,一个红色立方体,一个蓝色角锥,一个超级大方块。

当一个物体有了新的名字时,系统会使用新名字进行回答。

44. 谢谢。

不客气。

可以定义习语,但不必考虑语法。

注意到,SHRDLU 能够对不同的陈述、问题和命令作出反应;使用已有的知识;对输入作句法和语法的分析;解释和报告它的计划,并且具有对话策略(包括假定谈话者的目标和意图);作出符合常识习惯的应答。 116

5.3 SHRDLU 程序

SHRDLU 有十二个组成部分,如图 5.2 所示。威诺格拉德所作的最重要的突破是使这些组成部分能够取得相互联系。在计算过程中,加以某种约束的每个部分,都能按一定次序咨询其他组成部分。这部分地与人类行为类似,如 SHRDLU 在执行命令时,能将目标与问题解决结合在一起(上面

1—3 的对话),能够整合有关动词的信息(对话 13、14),并对问题作出回答(7—11和 15—29)。SHRDLU 通常在 15—20 秒内(1972 年的硬件水平)完成反应动作,“机器人”手臂的移动速度与人类的大体相同。我们最重要的几个目的是:①剖析它的语法结构;②了解它的语义集合程序;③了解它的认知—推理系统。

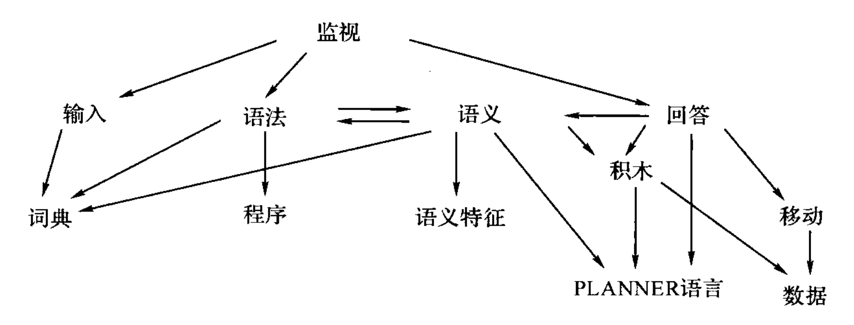


图 5.2　系统的组织结构(来源:Winograd,1972:5,图 1;经 Academic Press 许可复制)

句　法

句法上存在一些必然要解决的难题。例如,如何把每一个词语(名词、动词、形容词等)放到句子中的合适位置,然后再考虑词语和句子之间的语法(主语、谓语等)关系。例如,思考下面两个句子中“他”(him)的不同含义:

117

1. (a)亚瑟想要见他(him)(“他”不是亚瑟)

(b)亚瑟想要某人来看他(“他”是亚瑟)

SHRDLU 的句法是由称作 PROGRAMMAR 的一种特殊的以 Lisp 为基础的语言,对韩礼德(Halliday,1970)“系统语法”的实现。如图 5.3 是一个典型的分析句子的这种程序的例子。给定如下的一个句子:

2. 长颈鹿吃苹果和桃。

我们得到一个如图 5.4 的分析。句法分析并不是机械地和完全地进行操作;在分析的过程中,它会选择一条语义路径,并且在认知—推理系统中或者选择句法或者选择语义直接帮助分析。

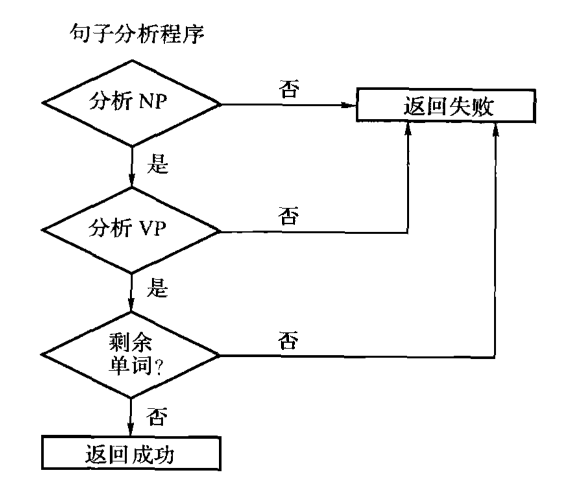


图 5.3　句子分析程序(来源：Winograd，1972：83，图 23)

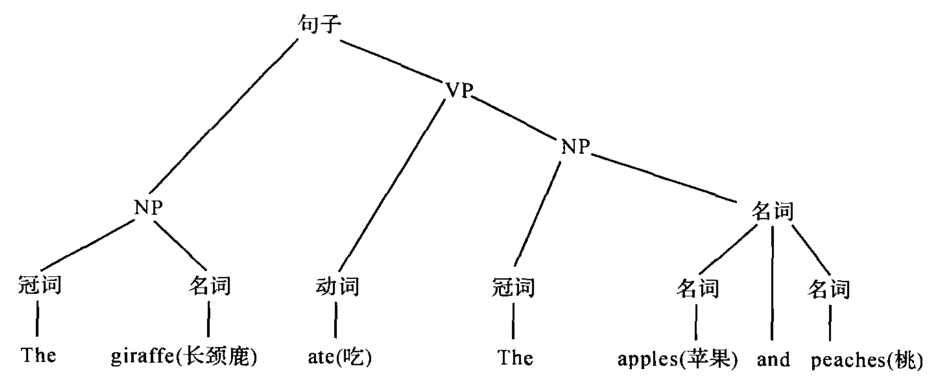


图 5.4　句子分析(来源：Winograd，1972：91，图 27，图 5.4—5.6；
经 Academic Press 许可复制)

语　义

SHRDLU 具有一个独特特征，称之为“过程语义”。语言表达的意义是由程序所做的事情来表征的，此程序由 PLANNER 语言的 micro-PLANNER 版本编写而成(参考 Hewitt，1971)：“这个模型潜在的基本观点是，所有语言的使用都可认为是以某种方式激活了听者脑内的某个程序。可以认为任何表达方式都是一种程序……”(Winograd，1973：170)。例如，118 “清理”这个动作的语义可以由程序 CLEARTOP(图 5.5)实现，扩展成如下一系列的句子：

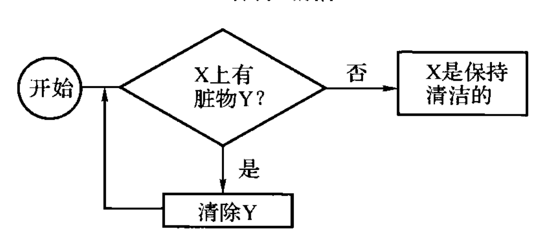


图 5.5　CLEARTOP 程序描述(来源:Winograd, 1973:169,图 4.5)

3.(a)将桌子擦干净(命令语气)
(b)能把桌子擦干净吗(祈使语气)
(c)桌子已经擦干净了(肯定语气)

每个句子都可以转化为一组 PLANNER 指令:直接的物理动作(命令),一些信息(问题),储存和修改一些当下的知识(状态)。如一个句子:

4.一个支撑角锥的红色立方体。

可由图 5.6 所示的 PLANNER 表示。

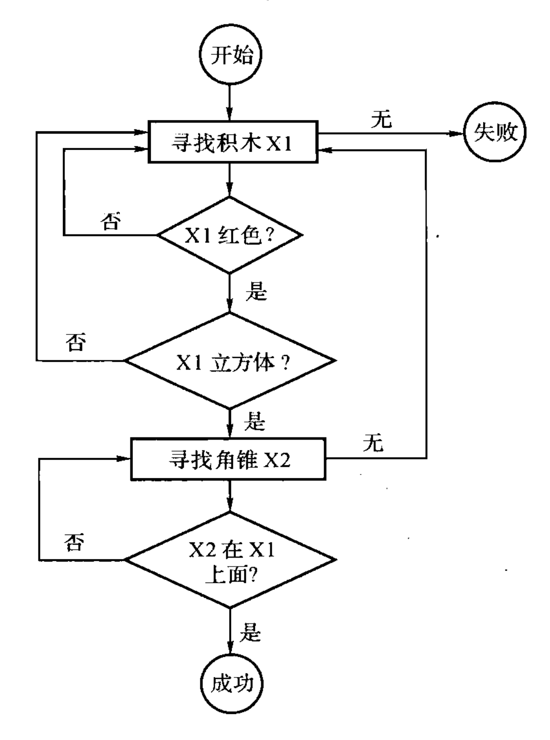


图 5.6　“支撑角锥的红色立方体”程序说明(来源:Winograd, 1973:172,图 4.7)

认知－推理系统

SHRDLU 的数据包含了如图 5.7 所列出的那些事实。这些事实加上 119 所有程序,共同构成一种一般推理系统,它通过运行一系列的子任务最终实现给定任务。现时运行的目标栈与图 5.8 所示相似。

```
(IS B1 BLOCK)
(IS B2 PYRAMID)
(AT B1 (LOCATION 100 100 0))
(SUPPORT B1 B2)
(CLEARTOP B2)
(MANIPULABLE B1)
(CONTAIN BOX1 B4)
(COLOR-OF B1 RED)
(SHAPE-OF B2 POINTED)
(IS BLUE COLOR)
(CAUSE EVENT27 EVENT29)
```

图 5.7　典型数据表达(来源:Winograd, 1973: 168, 表 4.1)

```
(GRASP B1)
  (GET-RID-OF B2)
    (PUTON B2 TABLE)
      (PUT B2 (453 201 0))
        (MOVED (553 301 100))
```

图 5.8　现时运行的目标栈(来源:Winograd, 1973: 171, 表 4.2;经 W. H. Freeman and Co. 许可复制)

5.4　SHRDLU 的局限

威诺格拉德(Winograd,1973)以非常谨慎的语言结束了他的讨论。他从如下两点,认识到了 SHRDLU 不足以作为人类自然语言程序模型。第一,程序是直接的:句法由正确的词组构成,语义(或者说认知一推理系统) 120 决定句子分析是否进行。但他认为这就使程序具有了过于严格的等级层次:"一种语言模型由某类在生物系统中发现的非严格的等级层次实现(如不同器官之间的协作一样),才能看起来像一种有效的心理理论"(1973:184)。第二,这个模型不能处理"在如两个智能人之间进行的交流过程中"(1973:184),他们共同默认的所有词语的含义。例如,对一些句子的理解之

所以有效，是因为听者知道说话者与他分享了一些关于真实世界的常识知识。思考下面句子中的"它"所指的含义：

5.(a)我将一瓶可乐掉到桌上然后**它**碎了。(瓶子或者桌子)
(b)拖把在哪：我将一瓶可乐掉到了桌上然后**它**碎了。(更倾向指瓶子)
(c)胶水在哪：我将一瓶可乐掉到了桌上然后**它**碎了。(更倾向指桌子)

121 如果程序能够模仿人类的这种能力，那么同样需要在程序内部构建很多常识知识。

后来，威诺格拉德(Winograd，1980：215)提出，要使SHRDLU与自然的人类语言程序更加类似，还存在另一些"显而易见的难题"。例如，也就是第三，由程序定义一个词语的概念，即使它比先前用逻辑形式定义的包含更多可能性，但仍然是不充分的。思考在"积木世界"几何形式的限制——角锥或者球体是由它的形状定义，不是通过它能做什么定义。同样，一个单身汉的定义是未婚、成年人和男性等属性，而不是他能够做什么。面对这样的问题，威诺格拉德为设计一种更好的知识表征系统，开始转向更加普遍且基础的问题(参见Bobrow and Winograd，1979)。

人工智能领域和认知科学内也发现了SHRDLU在一些基础方面的不足。例如，第四，SHRDLU对"积木世界"并不真正了解(颜色，有(owing)等等)。第五，对于如何"成比例扩大"SHRDLU或者关于这个微型世界的或者关于世界的知识，也毫无进展。第六，更重要的是，SHRDLU没有学习能力，虽然后来STRIPS程序在此方面得到了某些改进(Fikes and Nilsson，1971)。第七，SHRDLU并不真正理解语言的意义，它的"世界"并不是真实的。如福多(Fodor，1980a)写道："这种设计正是笛卡尔所担忧的那种情形：这仅仅是一台计算机，却偏要把自己想象成一个机器人"，回避处理和整合外部世界信息的难题。这个模型所具有的这些难题，我们后面还需要讨论。

5.5 SHRDLU的历史作用

尽管存在这些局限，我们也不应该忽视这样的事实，SHRDLU是运用程序将语义和句法分析与普遍知识进行整合的一次尝试。它表明，如果数据库足够狭小的话，那么它可以表现得更与人类的行为相似。这就激发了人们对专家系统的研究，如MYCIN(Shortliff)就已在工程学领域取得重要突破。

注释

[1] 这个命名,出自威诺格拉德与博登(Boden)间的一次私人交流(Boden,1977:501),"这些字母都包含在标准排字机键盘上的某一行内,排字机'校正'错误是通过把它们嵌入错误的一行文字内,这样校对员会很容易发现这个错误。在较差的校对过程中可能会导致这些错误都出现在印刷文章结尾处中——MAD的杂志中就经常出现这种情况。作为MAD的反对者,威诺格拉德用这些没有意义的字母作为程序的命名"。霍夫施塔特(Hofstadter,1979:628)也作了相同的解释。华尔兹(Waltz,1982:120)认为,"威诺格拉德的程序称作SHRDLU,就是从英文中出现频率最高的12个字母中挑选出7个组成"——加德纳(Gardner,1985:158)也持有相同看法。 122

[2] 引自Winograd(1972),后来Winograd(1973)又作了一些轻微改动。

【思考题】

ELIZA在短时间内能够进行智能对话,它是如何设计和实现的?
在SHRDLU中有没有什么例子能够在处理语言的输入问题上超过了被框定的回答?
SHRDLU在对话的能力上有哪些普遍特征?
SHRDLU包括哪五个部分?分别进行简述。
威诺格拉德与人工智能(和认知科学)界认为SHRDLU存在哪些局限?
SHRDLU在人工智能和认知科学中具有怎样的历史作用?

【推荐读物】

SHRDLU

理解SHRDLU最基本的运行情况,见Winograd(1972)。该书出版过一个缩略本Winograd(1973)。Winograd(1977)第2章介绍了将SHRDLU作为知识表征和理解人工自然语言计算机系统一部分的情形。Wilks(1977)第3章从一个完全不同的角度审视了这种情形。Winograd(1980)包含对SHRDLU的批判性探讨。Winograd(1983)是介绍计算语言程序的权威著作。Barr, Cohen and Feigenbaum(1981)第4章讨论了自然语言程序;第5章讨论了SHRDLU。新近比较流行的关于SHRDLU的讨论可以在McTeal(1987)中找到。Kobes(1990)说明了SHRDLU所喻示的虚拟方块世界在今天的认知科学界中产生的重要影响。

人工智能通史

Augarten(1984)很好地阐述了计算机发展的历史。从认知科学的角度研究人工智能的历史,可读性较强的是Gardner(1985)第6章。Minsky(1966)和 123

Waltz(1982)是学科领域内的研究成果，不过对于专业外读者(以及关注15年来人工智能发展和演变过程的人们)而言也具有较强的可读性。Barr, Cohen and Feigenbaum(1981)也总结和回顾了1980年以前的一些人工智能程序。Boden(1977)作出了一个很有影响的哲学取向的尝试。McCorduck(1979)是一部流传较广的通俗易懂的历史著作，Crevier(1993)也是如此。Dreyfus(1972/9)对人工智能的最主要观点作了回顾和批评——修订版(1979)将所关注问题的时间延伸到了1979年。Copeland(1993b)第1－5章回顾和评论了一些人工智能的主要研究项目。Feigenbaum and Feldman(1963)是历史上重要的且颇有影响的关于早期人工智能程序的论文选集。关于将来人工智能可能发展的历史(到2099年止!)，Kurzweil(1999)作了一些预言，在这本书中还介绍了一些有趣的与这个问题相关的URLs信息。

6 结　构

6.1　引言:基本概念

本章主要介绍和说明一些经典的数字计算组织或“结构”,并对其进行 124 分类。说计算机如何组织,就是说信息如何存储,以及什么决定了信息如何在系统中运作等等,并不关注计算机由何构成(继电器、真空管、转换器等),也不关注计算机运行了什么程序。有点像计算机的设计或蓝图,就像建筑的蓝图一样,并不关心材料细节和居住者的信息。但在我们开始研究计算机结构之前,首先需要弄清两组重要概念的区别。

算法/有效步骤与程序

我们首先需要对“算法”或“有效步骤”与“程序”作出严格区分。**算法**(或有效步骤)的本质是确保获得某一结果的一系列步骤。更精确地说,它是一组预先定义好的有限步骤序列,每一个步骤只占用有限的存储和时间完成,根据任何有限的输入得到某一结果。**程序**是指某种(编程)语言的有限指令列表。程序通常用于编码算法,可以认为程序能够运行算法。因此,任何运行程序的设备都能执行有限数目的指令,每个指令占用有限内存和有限时间,然后最终达到某一结果终止(实际上,它当然可以像文字处理器一样,等待其他指令的输入)。算法与程序之间的关系就像数(number)与数字(numerals)的关系。一个数可以由多种数字表示(如阿拉伯数字、罗马数字以及二进制数字),就像简单算法可以由不同的编程语言(如 Basic, 125 Pascal, Prolog)编写为多种不同程序一样。也可以这样认为,程序与算法之

间的关系可以比作词语(声音、形状)与其意义之间的关系——不同语言中不同的词语可以具有相同的意义(如 cat, katy, chat)。所以,某个程序有缺陷并不意味着算法有缺陷,很可能是编译它的语言出现了问题。

弱等价与强等价

我们还要区分两种算法"等价"的不同情形。说两种算法"弱等价"是指,给两个算法相同的输入会得到相同的输出。例如,有两种有关乘的算法,输入 32 * 17,都会输出 544。如果所有输入都会得到相同的输出,那么就可以说两种算法"**弱等价**"。但两个有关乘的算法得出结果的过程却可能不同。例如,第一种算法可能是 32 自身相加 17 次得到——可由一个加法器和一个计数器实现,这种算法称为**连续增加法**。第二种乘算法,可能如下:

```
   3 2
   1 7
 -----
 2 2 4
 3 2
 -----
 5 4 4
```

首先是 2 乘 7,向左进 1 位,然后 3 乘 7 加 1,得到 224。然后是 32 乘 1 得到 32,最后 224 百十位加上 32 得到 554,如上所示。我们注意到,这个方法首先是用 32 乘以 17 的一部分即 7,然后再乘以 17 的另一部分即 10,然后把这两个部分的结果加和得到最终结果,显然把这种方法称为**部分乘积加和法**是非常合适的。这种方法和连续增加法得出相同的结果——**弱等价**,即用不同方法得出相同答案。两种(或更多)算法**强等价**是指,不仅相同输入得到相同结果,而且它们的处理过程也是一样的,也就是说它们所有的中间步骤是一样的。显然,(1)如果一台机器运行的步骤另一台机器没有,那么可以直接说它们不是强等价。间接地,(2)也可从"复杂特性"中得出两台计算机的不同。例如,两台计算机都计算数量级较大的数的乘积,数字逐渐变大时,其中一台计算机处理的时间成比例增加,而另一台计算机处理时间并不成比例增加。或者从它们占用内存的大小或出现错误的次数判断两台机器的不同(参见 Pylyshyn,1979)。我们将会看到,强等价和弱等价的区别对于评估认知功能的计算模型在心理学上的可行性具有重要意义。我们想要得到某个模型,并不仅仅看它的认知功能能够做什么,还要看它的认知功能发挥作用的**方式**。心理实验就是典型地通过测量反应时和错误率,对某些认知功能及其模型的强等价测试。

126

6.2　图灵机

人物略传　阿兰·图灵(Alan Turing),1912年6月23日生于伦敦,曾就学于谢波恩中学和剑桥大学国王学院(学习数学)。1935年4月,冯·诺依曼从普林斯顿到剑桥访问,并在那里有一个月的授课活动,"在那个学期里,图灵当然认识了他,并且几乎从始至终参加了这门课程的学习……图灵可能就是这时听了冯·诺依曼课的原因,他在3月24日写往家里的信中说'我已经申请了明年去普林斯顿访问的基金'"(Hodges,1983:95)。图灵的确去了普林斯顿(于1938年获得数学博士学位),期间同阿朗索·丘奇(Alonzo Church)一起学习。到达一周后,他在写给母亲的信中说:"这里的数学系让我感到满意,这里有一大批杰出的数学家:冯·诺依曼、哈代(Hardy)、外尔(Weyl)、柯蓝特(Courant)、爱因斯坦。"1936年7月,他在普林斯顿发表了他最著名的学术论文《论可计算数及其在**判定难题**(Entscheidungsproblem)中的应用》[1]。在这篇论文里,图灵首次精确定义了"(自动)计算机器",即后来我们为了纪念他而命名的图灵机。图灵回到英国及随后的6年里,在外事局工作——实际上就是英国代码及加密学校,离伦敦50里远的高度机密机构。在这里,他的"Colossus"计算机帮助破解了德国"Enigma"密码,这项工作被认为缩短了欧洲战场两年时间。1945年,他辞去外事局的工作后,加入伦敦国际物理实验室,并领导设计自动计算引擎ACE(Automatic Computing Engine)项目。1949年,他接受了曼彻斯特大学的曼彻斯特自动数字机器"Madam"副主任职务。1950年,他发表了一生中最著名的非技术性论文《计算机器和智能》。在这篇文章里,他提出了"模仿游戏"的概念——即现在著名的用来测试智能的"图灵测试"。1952年,他因同性恋被判有罪。1954年6月7日,自杀身亡(他吃下了一个钾氰化物浸泡的苹果而死,也可能是意外死亡)。

127

图灵机(Turing Machine,TM)有相对简单的构造和操作模式,所以很容易用之证明,但却不可能用来实际计算,这就是为什么现实中无人制造它的原因。这里,我们首先了解图灵机的结构和运作方式,然后再介绍图灵机的一些基本特征。

结构和操作

直观上看，图灵机是由磁带和读写磁头（带有一个程序表和一个扫描器）两个主要部分构成的抽象计算设备。如图 6.1 所示，这幅图包含组成图灵机的四个主要部分：磁带、读写磁头、扫描器和程序表。

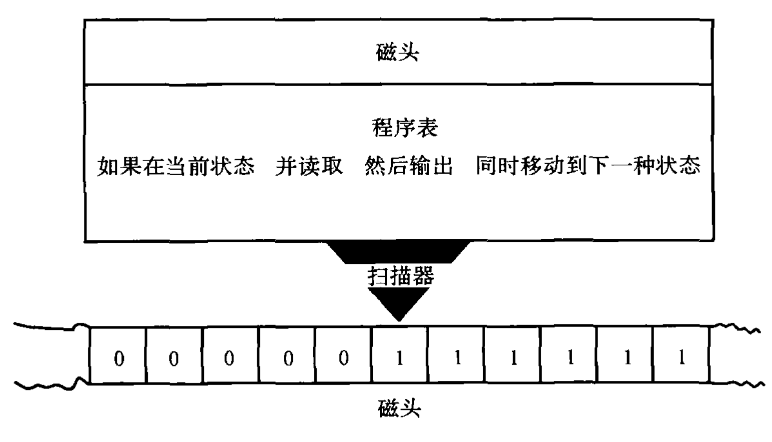


图 6.1　图灵机结构

结构

磁带：可双向无限补充，被分成一个个小格子，每个格子上要么是一个预先确定的字母符号，要么是空白。

读写磁头：读写磁头可以相对磁带前后运动，在磁带上读或写符号。

程序表（程序）：是一张有限指令列表，每个指令都包含具体的五种信息：（1）当前状态；（2）输入（读）字符；（3）输出（写）字符；（4）读写头向左或向右运动的命令；（5）下一种状态。

128　这些组成部分在计算过程中进行如下操作：

操作

控制：读写磁头的运动由程序表控制，而程序表主要关注两件事情：当前机器状态及其正在扫描的符号。得到程序表的指令后，读写头（1）在当前的小格子上写入特定符号，然后（2）向右或向左移动一个格子，或者保持在当前位置不变。然后，机器根据程序表的指令进入下一状态。

图灵机中运行的程序：包括具体的程序表和初始的磁带结构，当机器接收开始符号时便开始运行，遇到停止符时停止，磁带最后的结构状态包含着想要的输出结果。

计算过程：计算过程包括执行指令，由初始化的磁带结构和内部状态开始，再到遇到一个停止符结束。磁带最后的结构就是计算的输出结果。

输入：即初始的磁带结构形式。

输出：磁带最后的结构形式。

磁头

程序表

如果在当前状态	并读取	输出(P)	同时移动	到下一种状态
1	无	0	R	2
2	无		R	3
3	无	1	R	4
4	无		R	1

扫描器

图 6.2　初始图灵机(改自：Turing，1936/7：119)

图 6.2 是图灵第一台说明性计算机的示例。这台机器能够交替打印 0 与 1，中间用空格间隔。磁带开始时(在竖栏 2 为"无")为空白，机器要么打印"0"，要么打印"1"。"R(Right)"意味着机器将移到当下方格的右边方格，"P(Print)"意味着开始打印。例如，一个空白方格以状态 1 开始，然后：　129

状态操作

1. 打印一个"0"，然后向右移动一个方格，进入状态 2。
2. 向右移动一个方格，进入状态 3。
3. 打印一个"1"，然后向右移动一个方格，进入状态 4。
4. 向右移动一个方格(等等)。

图灵机循环：总结图灵机操作步骤，可以发现机器有一个循环：读、写、移动和进入下一状态。

已经证明了一些有关 TMs 的定理，但值得关注的是图灵最初提出的图灵定理(1936)：

图灵定理：存在一台“通用 TMs(UTMs)”，可以模仿任何其他 TM 的动作。

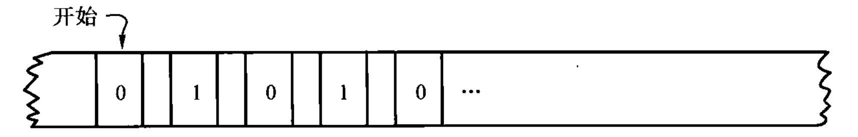


图 6.3　初始图灵机的操作

	1	2	3	4	5	6	7
Y	_L1	_L1	YL3	YL4	YR5	YR6	_R7
_	_L1	YR2	HALT	YR5	YL3	AL3	YR6
I	1L2	AR2	AL3	IL7	AL5	AR6	1R7
A	1L1	YR6	1L4	1L4	IR5	IR6	_R2

图 6.4　明斯基通用图灵机(来源：Haugeland, 1985: 139, 表 3;经 MIT Press 许可复制)

130　根据资料，已知的最小通用图灵机由明斯基设计(见图 6.4)。如果我们认为一个具体的 TM，能够通过编制或者形式化程序完成磁带上表征的事件，那么说通用 TM 能模仿任何具体 TM 的行为，就意味着通用 TM 能编码任何 TM 磁带上的描述。因为 TMs 是自动按步执行程序的，而 UTM 则能够清晰地编码有效步骤和算法的指令。这就引入了所谓的“丘奇-图灵论题”：

丘奇-图灵论题：对于任何可计算某一函数的有效步骤，都存在一个能够运行计算这个函数步骤的 TM(也就是说，做每一步骤所做)。

“有效步骤”是直觉概念，而不是一个形式概念，所以没有定理能够给予证明。然而，尽管每一次人们认为这个概念已经变得更精确了，新的概念还是被证明与图灵机等价。

术语小议 1931 年,科特·哥德尔(Kurt Gödel)提出了一组重要的**递归函数**,该函数可具体化为若干非常基础的数字操作步骤。1936 年,丘奇写道:“这篇论文的目的是提出一种有关有效计算的定义。”(参见 Davis,1965:90)丘奇运用的是“Lambda 定义”,并提出这种定义在他的“Lambda-演算”中与递归函数等价。他总结道:“两者相异且(在作者看来)具有相同有效计算的自然定义的事实,证明它们是等价的。这就增加了如下推论的强度,即它们构成了一种普遍的概念特征,正如概念通常由对概念的直觉理解构成一样”。此后不久,图灵在他的另一篇论文(1936/7)中认为,所有图灵机计算函数都是递归的,且所有递归函数都是图灵机可计算的。因此,有效计算可定义为 Lambda-定义,Lambda-定义与递归函数等价(丘奇),递归函数又与 TM 计算等价(图灵),因而有效性就是 TM 计算(丘奇-图灵论题):

131

丘奇(1936)
论题:有效步骤(直觉)=Lambda—定义
定理:Lambda—定义=递归函数

图灵(1936/7)
定理:TM 计算=递归函数

丘奇-图灵
定理:Lambda-定义=递归函数=TM 计算
论题:有效步骤=TM 计算

随后的研究发现,所有的各种形式化后能有效计算的系统(Normal 系统、Post 系统、Thue 系统、Semi-Thue 系统、Markoff 算法)都与图灵机等价。1957 年,王浩证明所有递归函数在操作程序图上都是可计算的(机器通过标准操作程序图说明计算)。最近,Fortran、Pascal 和 Lisp 程序语言证明也等价于图灵机。

由此可见,图灵机的计算与有效步骤弱等价。此论题的推论是,通用 TM 能计算任何有效步骤能实现的计算。因为我们的认知功能可视为有效步骤,UTM 能模拟这种功能的输入—输出对,所以它们弱等价。因此,在那段时期,有些人认为,这正为可通过恰当编程建造一台能够思维的机器找到了合理依据,并且认为可用脑内具有的有效步骤理解人类的认知功能。

6.3 冯·诺依曼机

132

人物略传 约翰·冯·诺依曼(John von Neumann),1903年12月28日生于布达佩斯,出生时起名为Neumann Janos (John) Lajos(按照匈牙利传统,姓放在前面)。因为他的父亲被授予了光荣的"Margittai(源自Margitta)"头衔,所以诺依曼的全名实际上为"Margittai Neumann Janos Lajos"。一家德国出版社将"Margittai"转译为德语"von",因此就变成了"John von Neumann"。1913年,小诺依曼被政府挑选接受国家教育,部分是因为他的语言天赋。10岁时,他就学会了法语、德语、拉丁语和古希腊语。童年时,他最喜欢学历史,通读了父亲总共44册的通史著作。1913年,10岁的他注册入读布达佩斯的一所著名高中,在那里接受了8年教育。1921年,进入布达佩斯大学学习数学。他习惯花大量的时间到柏林听哈伯(Haber,化学)和爱因斯坦(统计学原理)的讲课,只在考试时才会中断。1923年,迫于父亲的压力,接受"实能教育",进入了苏黎世ETH(Eidgenossische Technische Hochschule)学习化学工程。但他对数学的兴趣因经常与那时留居苏黎世的数学家外尔(外尔不在时,曾让诺依曼代课一个学期)和波利亚(Polya)的密切联系而保留。1926年,取得布达佩斯大学数学博士学位(他的学位论文后于1928年出版)。1927年,成为柏林大学无俸讲师,时年24岁——是这所大学历史上最年轻的教师。1930年,进入普林斯顿IAS(Institute for Advanced Studies)数学部。1943年访问英国时说道:"我开始对计算技术有了强烈兴趣。"由于他出色的流体动力学知识背景,应罗伯特·奥本海默(Robert Oppenheimer)的邀请,来到洛斯阿拉莫斯,作为一名爆聚(implosion)特性数学专家参加曼哈顿工程。他发现,模拟爆聚的计算问题,需要消耗大量的时间。所以在1944年,他用两周时间致力于改造机械操作穿孔卡和制表机连接板线路,"当他看到制表机连接板线路时特别沮丧,制表机在不同的计数器上平行操作,制表机连接板线路执行平行计算,需要考虑平行操作的相对时间。后来,**他的这段经历使他否定了电子计算机并行计算的可能,也否定了他的关于单个地址指令编码的设计,认为这种编码必然不能采用并行处理运算**"。

133

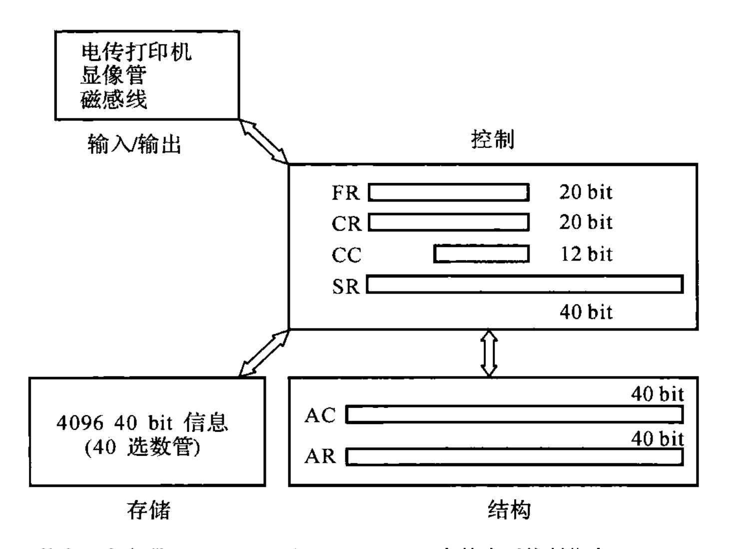
1944年9月7日,诺依曼访问莫尔电子工程学院(宾夕法尼亚大学),讨论了EDVAC(electronic discrete variable arithmetic computer)的设计构想,作为ENIAC的后继发展。1945年春天,在洛斯阿拉莫斯写就

了《EDVA 报告草案》,6 月份时他把这个初步草案邮寄到摩尔学院,此时《报告草案》还有一些引文仍留有空白,没有标题。摩尔学院为之添加了一张封面就此出版,诺依曼的这篇尚未完整的报告草案就这样流传开来。"文献非常有力地,仅用 100 页的油印文本就给定了程序存储计算机的基础……没过多久,美国及英国出现了很多对建造这种高速计算装置感兴趣的不同组织团体。事实上,他的工作,为很多早期程序储存计算机提供了一种逻辑程式"。1955 年,诺依曼被诊断患有骨癌。1956 年,他为斯蒂尔曼讲座(耶鲁大学)写作了《计算机和脑》。1957 年 2 月 8 日离世。

"冯·诺依曼机(vNMs)"或者称之为"寄存器结构"机的设计,部分地克服了图灵机的不足[2]。冯·诺依曼机的原型是依据冯·诺依曼最初关于 EDVAC 的研究(1945)建造的,成为勃克斯、哥德斯坦和诺依曼在普林斯顿高等研究院(IAS)关于计算机储存器研究的基础(Burks, Goldstine, Neumann, 1946:98—9):

> 1.1 *[完全自动]* 这个设备完全是通用目的的计算机器。包含一些元件,涉及算法、存储器、控制器及与人类操作的连接。这表明,计算机具有完全自动的特征,如当计算开启后可不依靠人类操作而独立完成……
>
> 1.2 *[储存器]* 显然,机器不仅需要对特定计算需要的数字信息以某种方式储存……而且对控制实际执行数字数据路径的指令也能够存储…因此必须有某种元件,能够存储这些程序命令。而且,还需要有一个单元能够理解并执行这些指令。
>
> 1.3 *[程序存储]* 上面我们已经概括地讨论了两种不同形式的存储:数字存储和指令存储。如果把对机器的指令化简为数字编码,并且如果计算机能以某种方式区分数字和指令,那么储存元件就能同时储存数字和指令……
>
> 1.4 *[控制器]* 如果存储元件仅仅是对指令的存储,那么还需要存在一个能够自动执行存储于储存器内指令的元件,称之为"控制器"。 134
>
> 1.5 *[运算器]* 因为装置是计算机器,所以必须存在运算元件,它能够进行一些基本运算操作……
>
> 被看作是机器基础的这些操作,通过机器线路实现……
>
> 1.6 *[输入—输出]* 最后,必须存在的设置是输入和输出元件,这样操作者才能与机器互相交流……

这些部分的构成关系,可由图 6.5 表示。派利夏恩(Pylyshyn,1984)对冯·诺依曼的最初设计说明作过一个合理的评论:"事实上,人们对每一个广泛有效的结构都会认为是冯·诺依曼提出的某一种类型(虽然使用这个135 名字常常意味着是指'传统计算机')。这种结构——从普林斯顿 IAS 计算机设计开始得到了广泛应用——是一种寄存机器,通过数字'地址'储存及检索符号,通过程序的序列转换(除分支指令)实现控制,符号操作则由如下步骤完成:先从存储器中检索符号,把它们储存在指定的寄存器里,再提供一个原初命令,然后再将结果符号重新储存在存储器内。虽然存在各种变式,但自从数字计算机开始,**通过一系列'提取'、'操作'和'存储'控制而实现序列加工的主要思想,一直占据支配地位**"(Pylyshyn,1984:96-7)。



FR:函数桌面寄存器(Function table register)——存储当下控制指令
CR:控制寄存器(Control register)——存储下一条控制指令
CC:控制计数器(Control counter)——存储下一条控制指令的地址
SR:提取寄存器(Selection register)——存储从内存提取的数据或将数据发送到内存
AC:计算器(Accumulator)——用于算法操作
AR:算法寄存器(Arithmetic register)——用于连接 AC

图 6.5 冯·诺依曼机结构(来源:Pohl and Shaw, 1981: 141, 图 5.1; 经 W. H. Freeman and Co. 许可复制)

冯·诺依曼机循环:总结冯·诺依曼机操作循环如下:提取、操作、存储。然而,每次循环只可执行一个指令,这就对计算机信息流量产生了约束。这种约束有时称为"冯·诺依曼瓶颈"。

冯·诺依曼机的程序运作：物理句法

我们已经知道，冯·诺依曼机的基本操作循环是提取指令、操作（执行）指令和存储结果。程序是使用某种编程语言执行某种算法的一组指令。我们现在要介绍的是，至少是简述，程序怎样在冯·诺依曼机上运作——我们怎样使一台物理机器能够做我们告诉它做的事情？或者更夸张地说，如果程序是计算机的“心灵”，触发器、寄存器以及电路闸是它的“身体”，那么我们如何解释计算机的心一身问题呢？[3]

我们现在以简单的汇编语言程序比较两个内存地址的内容是否相同（寄存器 1 和寄存器 2），来说明这个问题。如果它们相同，计算机显示“OK”，如果不相同则显示“NO”。下面就是这个程序（选自 Copeland 1993b, chapter 4）。大括号里的内容是对每个指令的说明。

“比较和输出”程序：

1. 比较：寄存器 1 和寄存器 2

{比较寄存器 1 和寄存器 2 的内容，如果它们不同，将 0 放入匹配寄存器内，如果相同则放置 1}

2. 分支——0 6

136

{检查匹配寄存器，如果是 0，则跳至第 6 行}

3. 输出——字符 1001111

{这是字母 O 的 ASCII 码}

4. 输出——字符 1001011

{这是字母 K 的 ASCII 码}

5. 分支 8

{跳至第 8 行}

6. 输出——字符 1001110

{这是字母 N 的 ASCII 码}

7. 输出——字符 1001111

{这是字母 O 的 ASCII 码}

8. 停止

程序说明

从第 1 行开始，程序检测寄存器 1 和寄存器 2 的内容，对之进行比较。

如果相同则存储“1”，如果不同则存储“0”。在第2行时，如果内容不匹配，程序则跳至第6行，首先输出N，然后输出O，在第8行停止。如果在第2行时，两个寄存器匹配，则相续输入O和K，即OK。所以，如果数字不匹配，计算机回答NO，如果匹配则回答OK。作出任何一种回答之后，机器即停止。

程序编译

因为这个计算机是二进制数字机器，所以所有的指令都要转换为二进制编码。将指令转换成二进制编码的工作由编译器完成，编译器还负责访问寄存器和对数字行线的编译。在这个程序里，程序的一个简易bit码列表，可能如下：

操作	Bit-code
比较	00000011
分支0	00000100
分支	00000110
输出字符	00000101
停止	00000001

有关寄存器中的指令，也必须使用的二进制编码，这样计算机才能找到它们。

137

寄存器	Bit-code
1	01
2	10
21	10101
25	110111

程序编译结果

汇编语言编译程序的结果，可能如下：

指令	寄存器/字符
1. 比较 00000011	1 10(寄存器:1,2)
2. 分支 0 00000100	10101(寄存器 6)
3. 输出字符 00000101	1001111(O)
4. 输出字符 00000101	1001011(K)
5. 分支 00000110	1100111(寄存器 25)

6. 输出字符 00000101　　1001110(N)

7. 输出字符 00000101　　1001111(O)

8. 停止 00000001

编译器将这些指令和寄存器分配到储存器的地址中，这个操作使获得了一个地址寄存器(例如，下文程序编译结果中第 1 行中的“10”)，寄存器获得另一名称(例如，下文程序编译结果中第 1 行中的“11”、“12”)。

储存器中的程序编译结果

储存器中的程序编译结果，可能如下：[4]

第 1 行 {比较寄存器 01,10}

10　　00000011

11　　1

12　　10

第 2 行{如果 0 跳至寄存器 10101＝21}

13　　00000100

14　　10101

第 3 行{输出字符 1001111＝O}

15　　00000101

16　　1001111

第 4 行{输出字符 1001011＝K}

17　　00000101

18　　1001011

第 5 行{跳至寄存器 1101＝25}

138

19　　00000101

20　　1101

第 6 行{输出字符 1001110＝N}

21　　00000101

22　　1001110

第 7 行{输出字符 1001111＝O}

23　　00000101

24　　1001111

第 8 行{停止}

25　　00000001

每个寄存器是一列触发器，每个触发器只能出现两种状态中的一种状态：或开，或关。所以，编译程序的计算机等同于某一特定触发器构造——有些为开，有些为关。为运行程序，需要人按下一个按钮，或者等价地通过将第一个指令复制到指令寄存器中，开启物理进程。如此设计的计算机（硬连线），当指令寄存器的触发器设置成如 00000011 时，它便可执行一个比较过程——将寄存器 1 和 2 的内容进行比较，在匹配寄存器里产生 1 或者 0。程序展示的下一个指令被复制到指令寄存器里，机器持续这样的提取－操作－存储循环直到停止——符合所有的物理以及电子工程法则。机器里没有虚幻的魂魄，只是程序指令的句法编译成物理触发器的不同排列。这就是我们是怎样使机器做我们命令它做事的。

冯·诺依曼机与图灵机

冯·诺依曼机克服了图灵机的许多缺点：(1)NMs 的存储既允许直接（完全的、随机的）存取，也可以间接（或相对的）存取，而 TM 只允许间接（或相对的）存取。(2)程序作为数据也可在存储器中储存。(3)具有现实的计算元件作为算法单元（TMs 没有现实的目标的环路）。有人认为，前两点是“冯·诺依曼机结构的重要突破”(Haugeland，1985：143)。冯·诺依曼结构允许在全部过程中插入“子路径”，也就是说，可在计算过程中的任意一点进行调用。当子路径执行结束后，计算还会从它断掉的地方开始，就好像这个子路径只是程序的前指令。子路径的使用引起了计算系统的“模块理论”，这意味着一旦子路径能够正确运行，可以插入到程序中任何需要的地方完成运算。且不管被调用多少次，只储存一次。如果一个程序，由很多之前已经调试过的子路径组装而成，那么它运作起来功能会更好。如果有些地方出现了错误，子路径可以较快地分离出这个错误点：首先找到错误子路径，然后处理。最后，是机器**如何**计算它所计算的函数的问题。不同的结构有不同的“复杂特性”——运行不同算法所需步骤数量、时间以及使用内存的大小都不相同。在个别情况下，这两种不同的结构并不能直接执行相同的算法，“例如，用一台图灵机检查一个符号串需要的基本步骤数，**随所储存的符号数目的平方而增长**。而如果用‘寄存器结构’（通常称作随机存取存储的加工结构…）（指冯·诺依曼机）来完成这件事，在特定条件下其时间复杂性**与所储存串数目无关**……有些事情图灵机是不可能完成的，尽管在算法上图灵机与之弱等价……而寄存机器……却可适用于多种算法，包括二值搜索。在二值搜索过程中，寄存机器运用比较而将剩余选项组的筛选范

139

围逐渐缩小，如'20 问题游戏'[①]……这在图灵机的计算结构中**不能直接执行**"(Pylyshyn，1984：97-9)。

6.4　产生式系统

纽厄尔和西蒙(Newell and Simon，1972；Newell，1973)提出的产生式系统(Product Systems，简称 PSs)沿用了波斯特(Post，1943)系统相同的名字。产生式系统在**心理学**中广泛应用于模拟各种认知功能，在**人工智能**领域常常用于构建专家系统。

纽厄尔和西蒙(1972)明确说明，PSs 用来模拟认知现象，与行为主义关注刺激和反应关系有很大区别(见第 2 章)，同样区别于那个时期神经科学家所关注的神经硬件问题(见第 3 章)。他们写道："本书讨论的信息加工理论，代表一种处于行为主义和神经科学之间的一种特殊解释层次"(1972：876)。纽厄尔这样描述 PSs 的特征："产生式系统是一种具体的信息加工系统程式。它包含一组产生式集合，每个产生式都具有一个条件和一个动作。它还具有数据结构库：编码产生式系统运行信息的表达式集——是动作的操作对象，并决定条件为真或假。当给定初始结构数据后，产生式系统进行如下操作：如果当前数据条件为真(假设当下只存在一条数据)，产生式则被执行。这会使动作对当前的数据结构产生修改。修改后的数据结构会紧接着引起另一个(可能会是与上一个产生式相同的)产生式被执行，数据结构进一步修改。每个动作操作均由当前结构数据库的条件为真引发。一个动作操作紧接着另一个动作操作运行，这样持续进行，也就执行了整个加工程序。如果整个加工停止，则只有两种可能：要么没有条件为真(因此什么都不会引发)，要么结论本身就包含一个停止操作指令"(1973：463)。纽厄尔所述的产生式系统结构，如图 6.6 所示。

140

---

① 20 问题游戏(Twenty Questions Game)，参加者包括回答者和提问者。首先，回答者在心里面先想好某一确定对象，如北极熊。然后，提问者开始对回答者提出回答者只能作出"是"、"不是"或者"不知道"回答的问题。例如，问"是否是一种动物"，答"是"。回答者不能说谎，是游戏得以进行的前提。提问者在提出 20 个问题之前，如果猜对回答者心里想的对象目标，则赢得游戏。按照信息理论，此游戏最多可确定有关某一对象的 20 bit 信息。也就是说，如果每一问题可消除有关某一对象的一半不确定性，那么 20 个问题可区分 2^{20} 种情形。因此，提问者的最佳提问方式是使每一问题能对所有剩余可能性中的一半得到确认或消除。文中说的"二值搜索"，就是采用这种策略，将"剩余选项组的筛选范围逐渐缩小"，最终得出最佳答案。——译者注

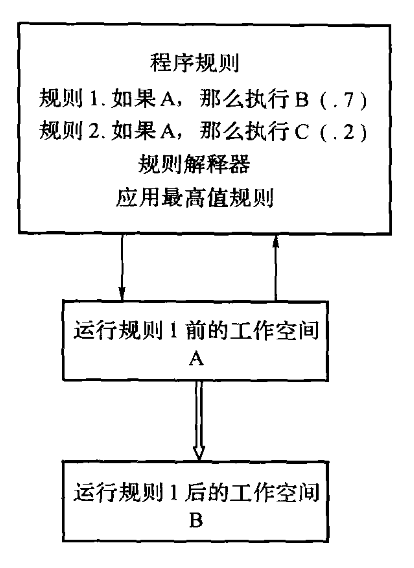


图 6.6 产生式系统的构成和操作(1)

PSs 包括三个组成部分。**产生式规则集**(a set of production rules)，它的形式是：如果 A(条件)，则 B(动作)；**储存工作空间**(memory work space)(有时称"语境")；**规则解释器**(rule interpreter)，它将相关的规则应用于工作空间的结果。PSs 的典型**操作循环**：

141

1. 把生产式的条件(A)与工作空间的结果相匹配。
2. 从产生式间的冲突进行消解，并将正确的产生式写入工作空间。
3. 在工作空间中执行 B。

产生式系统循环

总结产生式系统(PS)的操作循环为匹配、消解以及运行。

当不再有产生式适合工作空间时，系统就会停止(或者遇到特殊的停止符)。产生式通达工作空间的竞争是并行的——只有激活了的产生式才是序列的，这一点图 6.6 显示得并不是很明显，图 6.7 可以更好地加以说明。

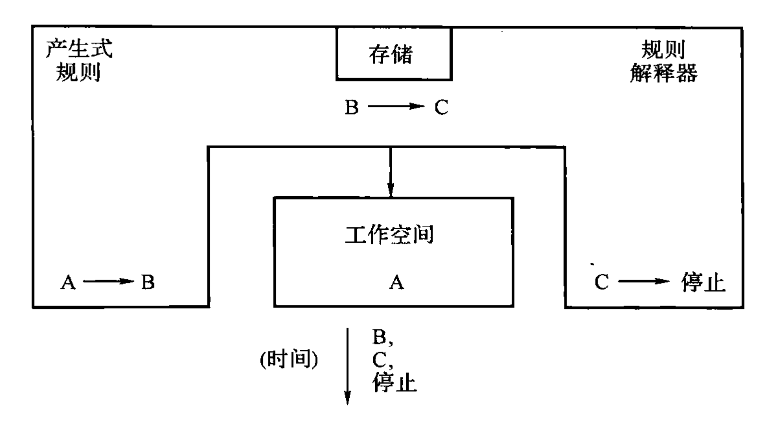


图 6.7　产生式系统的构成和操作(2)

可以通过编写产生式规则集对产生式系统进行**编程**，产生式规则在工作空间输入初始数据后开始运行，当工作空间产生目标输出结果后停止。通过变换不同的工作空间和使用时间，可以得到各种特殊的产生式系统(参见 Newell，1973：516)。例如，图 6.8 是一个简易产生式系统的示例。

142

产生式规则集：

P1. IF (On-CL 绿色) THEN (Put-On-CL 农产品)
P2. IF (On-CL 装进一个小容器内) THEN (Put-On-CL 美味)
P3. IF (On-CL 冷冻 OR On-CL 农产品) THEN (put-On-CL 易坏)
P4. IF (On-CL 重 15 磅 AND On-CL 便宜的 AND 并非 On-CL 易坏)
THEN (Put-On-CL 常用的)
P5. IF (On-CL 易坏 AND On-CL 重 15 磅) THEN (Pub-On-CL 火鸡)
P6. IF (On-CL 重 15 磅 AND On-CL 农产品) THEN (Put-On-CL 西瓜)

规则解释器：

1. 查找所有条件部分为真的，可用产生式。
2. 如果可用产生式多于一个，那么执行所有的产生式，复制其动作符号加入 CL。
3. 执行最低限数的(或者只有一个的)可用产生式的动作。如果无产生式可用，停止。
4. 重置所有产生式的可用性，并返回 S1。

图 6.8　简易产生式系统(来源：Barr，Cohen，and Feigenbaum，1981：191)

将产生式应用工作空间(这里称为“CL”)中的数据，得到如下计算：

1. ON-CL＝绿，重 15 磅

{在工作空间中输入数据：绿，重 15 磅}

P1　激活

2. ON-CL＝**农产品**，绿，重 15 磅；

{P1 将“农产品”加入工作空间}

P3　激活

3. On-CL=**易坏**,农产品,绿色,重 15 磅

{P1 将"易坏"加入工作空间}

P5　激活

4. On-CL=**火鸡**,易坏,农产品,绿色,重 15 磅

{P5 将"火鸡"加入工作空间。不清楚为什么 P5 不被禁止——将"非农产品"加入它的条件}

P6　激活

5. On-CL=**西瓜**,火鸡,易坏,农产品,绿色,重 15 磅

{P6 将"西瓜"加入工作空间。不清楚西瓜和火鸡是什么}

6. HALT

这个例子说明了编译 PSs 程序存在的一些困难。这个例子中的错误提示是,需要对产生式精确地编写,这样才能正确地使用。注意到,如果 P5 和 P6 的指令得到修正,那么"西瓜"就成为最后的条件,这样就不会出现错误了。

143

在工作空间中,有多种对于相互冲突产生式的消解方法。例如,可在匹配过程中增加一种测量方法,确定哪一个最为合适。或者,根据每一个产生式先前的激活情况,使最活跃的(或依据频率,或依据新近发生)产生式发生激活(有关讨论和检测,参见 Anderson,1983)。

产生式系统与冯·诺依曼机

值得注意的是,产生式系统的结构有很多不同于 NMs 的地方,通常被认为是其**优点**。首先,与 TMs 和 Neumanns 机相比,除了解释器,产生式系统是非分离的、外部的控制结构;除了冲突消解,有且只有条件与工作空间的内容相匹配,产生式才会被激活。这就使系统具有了**并行加工**的特点,但正如我们看到的,这同样也使 PSs 对具体的系列动作的反应,出现了困难:机器规则,对于想要得到的目标算法结构很难满足(我们还将在很多联结主义结构中遇到类似的问题)。第二,PSs 依据工作空间对信息的描述进行操作,而非依据完全或相对地址——它是"内容"寻址而非"定位"寻址。试想一下,警察到某地(匿名端)逮捕那里的所有人(定位寻址),和寻找与犯罪现场留有的指纹相匹配的人这两种抓捕罪犯方式的不同。其实,冯·诺依曼机也可通过 Hash 编码方式实现内容寻址。内容定位处,或者说内容本身,是输入参数和探测器的逻辑一运算函数[5]。然而,这些技术只对信息的强

约束形式有效。甚至,即使对信息发生作用,这些方法也需要通过"碰撞函数"避免 Hash 编码产生多重或错误的地址。这就使冯·诺依曼机在降解过程(degradation)中变得非常不稳定和繁琐。而 PSs 通过运用匹配一子循环的测试使降解过程更为通畅。第三,有学者认为 PSs 比之 vNMs 更具**强模块化**优势,"因为产生式系统具有强模块化优势,那么必然存在一种统一地使之得到扩充,而不会使分布变更扩展至整个现有系统的方法"(Pylyshyn,1979:82)。豪格兰德(Haugeland,1985:162)走得更远,他说:"一定程度上,生产式系统的'模块性'是其他任何结构无法比拟的。**模块**就是某一能完成某项完备定义任务的独立子系统,并且它能以简约的方式与系统内其他子系统相互交流……产生式系统……当条件满足时就自行激活,不论其他产生式所知或者其决定条件是什么"。PS 中的产生式类似于 vNM 中的指令,但产生式的增加或消减(例如,在程序某处需要作出些许变更)比之诺依曼指令更为自由——具有冲突消解功能。但是,我们将看到 PSs 的这种模块性优势也令其具有很多缺点。 144

PSs 的**缺陷**主要有:(1)**效率低**(既然它需要执行每一个满足条件的动作,那么就很难快速地完成预先确定的一系列步骤);(2)**模糊性**(很难使之实现表达或建构算法结构)。这两个缺陷主要源于产生式的模块性和统一性——不能通过子路径递层调用其他产生式,而这一点诺依曼程序是可实现的。

6.5 魔域混战场:鬼蜮模型

最后我们要介绍一种机器——鬼蜮模型(Pandemonium),与其把它看作是一种普遍计算结构,不如说它是关于模式识别与学习的一种模型。尽管这种模型是在联结主义兴起的多年以前提出的,但它却包含了很多当代联结主义的普遍观点,同时它也是基于与联结主义相同的思考而产生的。因此,鬼蜮模型同时具有数字与联结主义模型的部分特征。塞尔弗里奇(Selfridge)说道:"这种模型背后的主旨是并行加工。这是因为:首先,使用并行方式处理数据比较容易,并且也更为贴近真实的'自然'处理方式;其次,相对独立的准模块集合,比起所有部分都是以某种复杂的方式直接联结起来的机器来说,出现错误时更容易得到修正"(1959:513)。因此,"鬼蜮模型……似乎并不像以前众多观点认为的,具有各种固有限制或缺乏弹性"(1959:513)。

结 构

在简化的鬼蜮模型中，包含一个判断“鬼”（我们知道，模型的名字毕竟是 PanDEMONium）的部分，它负责监听所有的认知“鬼”。每个认知鬼都注视着输入数据，并试着识别与它专有特征或属性相近的数据。数据越接近，145 认知鬼的喊叫声越大，判断鬼便会对喊声最响的那个认知鬼进行选择。图 6.9 是这一过程的简化示例。

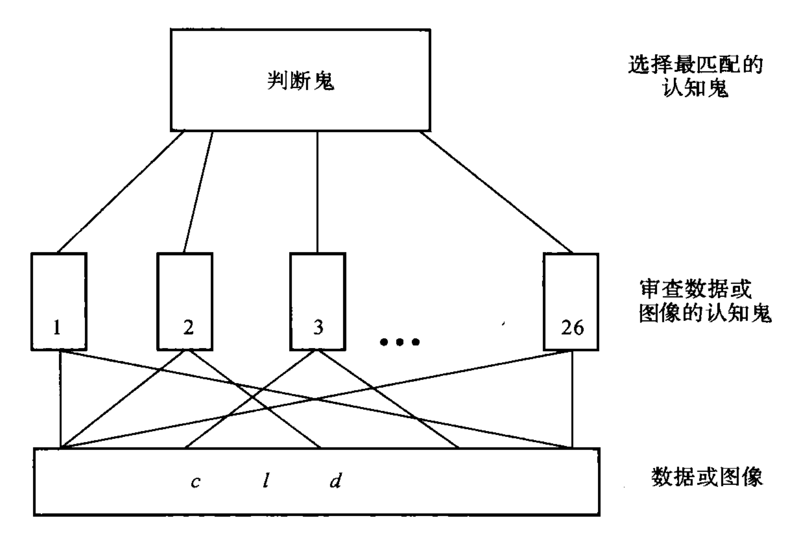


图 6.9　理想化的鬼蜮模型（来源：Selfridge，1959：315，图 1）

例如，如果数据是字母表中的一些字母，1—26 只认知鬼，每只鬼只专注于其中的一个字母。因此，尽管输入的 *d*，*c* 和 *l* 都相似，但它毕竟是 *d*，所以 *d*-鬼喊得会最响，“很多时候，一种模式等价于某组特征的逻辑函数，这组特征中每一个都可能是很多其他模式的组成部分，它所缺失的特征其他模式可能也缺失”（Selfridge，1959：516）。所以，塞尔弗里奇对理想化的鬼蜮模型作了修正，增加了计算鬼层（见图 6.10）。这样，认知鬼需要加上计算鬼的加权和。计算鬼层独立于目标，必须经过适当的设计（通过自身演化或实验者安排）才可使之能够学习。

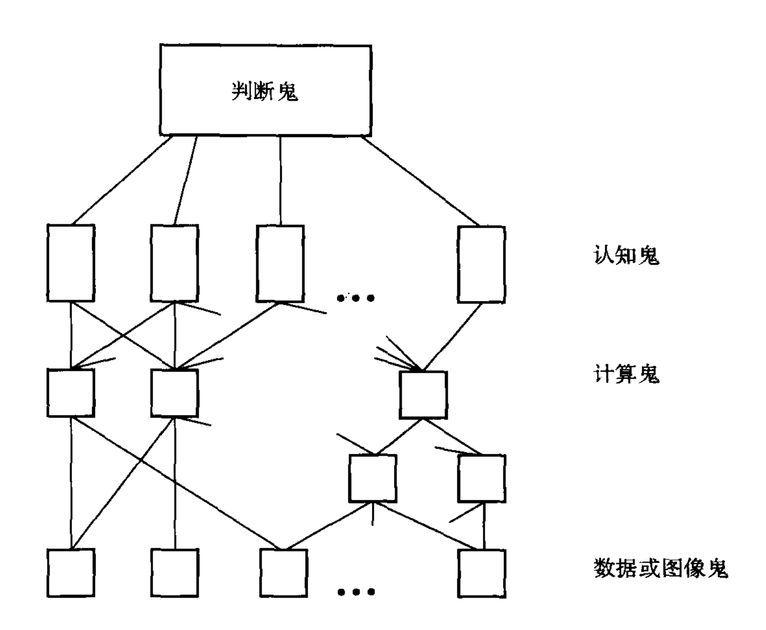


图 6.10　修正的鬼蜮模型(来源:Selfridge,1959:517,图 3)

学　习

鬼蜮模型这样的设计就是为了使它能够学习。塞尔弗里奇描述了两种鬼蜮模型的学习程序。第一种是通过“特征加权(feature weighting)”进行学习,“认知鬼为使在整个鬼蜮模型中的得分最高,可修正指派到次鬼层的权值”(1959:518)。之后,塞尔弗里奇还讨论了几种修正特征一权值规则的可行方法。第二种学习程序是“次鬼选择(subdemon selection)”,“第一种程序采用特征一权值适当修正的策略使权值得到优化,但我们并不能保证那些选择的次鬼是适当的。次鬼选择能够产生新的竞争次鬼用以替换效率低下的次鬼,也就是那些对提高分数没有太大帮助的次鬼”(1959:521)。塞尔弗里奇设想,控制同样受鬼本身变化的支配,“原则上,我们认为控制操作自身要服从于鬼,如同特征一权值和次鬼选择,受其本身变化的支配”(1959:523)。最后,塞尔弗里奇认为,到目前为止,他对鬼蜮模型的描述“都需要以人的持续监控为基础,当机器出错时必须及时告知机器”(1959:523)。但我们或许想知道,机器是否能够凭借自身提高它的表现(学习)能力呢?“我认为可以这样做到,即模型必须具备一种极其明确的正误判断标准,有且仅有

146

一个认知鬼的输出比所有其他的认知鬼更符合目标结果”(1959:523)。塞147尔弗里奇简述了鬼蜮模型机器对莫尔斯码的识别，或者更精确地说是“它能够区分，以手动键盘输入的莫尔斯码的点和线”(1959:524)，但他并没有对此作进一步的解释。

塞尔弗里奇和奈瑟(Selfridges and Neisser,1960)报告了 CENSUS 的最初实验结果。CENSUS 是一种具有鬼蜮模型结构的机器，能够识别 10 个手写字母：A,E,I,L,M,N,O,R,S,T。输入被投射到 32×32(1024)像素的“图片”(输入储存)上。CENSUS 能够应用 28 个特征来计算图片上显示的是哪一个字母的概率。根据塞尔弗里奇和奈瑟的研究，CENSU 的识别正确率只比人类低 10%。但若要进一步扩展机器的功能，至少还存在三个重要难题：首先，如何从潦草的文字中识别出不同字母；其次，如何运用学习提高其判断正确率；第三，如何利用学习获得其自身的特性。它现有的这些功能还都仅限于程序设计者的预想范围之内。如，他们在文章的结束语中写道：“我们仅仅是在猜想，这种限制如何可能得以克服。除非如此，‘人工智能’终将会被误认为不会具有独立智能”(1960:68)。甚至 20 年以后，联结主义到来的时代——20 世纪 80 年代早期，这种误解始终没有消除。

鬼蜮模型、感知器和联结主义

我们容易发现，鬼蜮模型对于联结主义的发展所起到的重要作用。鬼蜮模型与感知器一样，都采用高度互相联结的并行加工单元从而实现模式识别。尽管每个单元都仅处理少量的信息，但是通过局部的相互作用，能在整体上作出正确的判断。而且，与感知器一样，系统也通过使用各种学习方法使联结权值发生变化，从而能够进行学习。也许鬼蜮模型的独特之处，在于控制鬼和它们如何进行选择的假设。在感知器或当代联结主义理论中与这些概念存在哪些相似之处，尚不完全清楚。

6.6 结构分类(Ⅰ)

一种建构分类的方法是依据某些通用原则为基础，先判断哪些范畴可能会被用到，然后对各个系统进行考察，之后找出它们最符合哪种范畴。这148种方法存在的风险是，可能并没有按照它们本身内在的某种维度进行划分。另一种建构分类的方法是先对样本(来自不同类别)进行观察，然后尝试概

括范畴的维度。这种方法的风险是，可能会从数据中得出某些互斥的推论。我们前面已经考察了一些典型标准结构，是否存在一种根本性的对之进行分类的方法呢？显而易见，对于一种(数字)计算设备来说，至少全部包含三个重要部分，即表征、存储和控制：任何(数字)计算设备必须能够**储存表征**，并且都必须能够通过某种具体方法对状态进行转换——**控制**。因此，这三个维度可作为结构划分的基础。我们现在暂时先不考虑表征，只对存储和控制进行总结(见图 6.11)。

	图灵机	冯·诺依曼机	产生式系统	鬼蜮模型
控制	定位式	定位式	部分分布式	分布式?
存储	间接	直接定位寻址	直接内容寻址	直接内容寻址

图 6.11　结构类别

如果机器的所有控制都只源于唯一部位的操作，控制称为**定位式**，否则则为**分布式**。如果机器的存储是组织化的，想要寻找一个地址必须先经过另一个地址，那么说它的存储是**间接的**，否则为**直接的**(或随机的)。如果数据的储存和读取只在某固定地址中，那么说存储是**定位寻址**；如果数据依据内容进行储存和读取，那么存储就是**内容寻址**。

由于这些结构特征的原因，对应机器的计算也就有了某些限制。需要强调的是，图灵机和冯·诺依曼机的计算是串行的，而 PSs 是部分并行，鬼蜮模型则几乎全部为并行计算。

注释

[1] “判定难题”是指存在着哪些一般步骤，可用于判断某些系统内的某一任意公式是否为系统公理。

[2] 我们将这些机器称之为“冯·诺依曼机”，是因为这种结构的机器起源于冯·诺依曼 1945 年的 EDVAC 报告。虽然图灵(Turing，1946)也因为他的一些重要观点赢得此项荣誉，但是我们不能同时有两种图灵机。

[3] 心一身问题，即心理现象与生理现象关系的问题，后文还会继续阐述。

[4] 为方便理解，左列数字是采用十进制表示的寄存器序号(计算机中的地址)。　149

[5] 例如，输入参数中间的 2 bit 信息可表示其内容的地址，或者将异或逻辑用于两个半个字符串，或者将输入字符串截断为两个等长的字符串，然后进行算术加和。

【思考题】

引言

什么是算法/有效步骤与程序?

什么是计算机的"结构"?

两种机器的强等价与弱等价分别指什么?

图灵机(TMs)

图灵机的主要构成是什么?

它们是如何组织的?

图灵机是如何操作的?

图灵机如何编程?

描述图灵机的输入、输出以及计算过程。

图灵机循环指什么?

什么是图灵定理?

什么是丘奇-图灵论题?

为什么丘奇-图灵论题是论题而不是定理?

把定理和论题放在一起对认知科学有什么影响?

冯·诺依曼机(vNMs)

vNM 的主要构成是什么?

它们是如何组织的?

vNM 循环指什么?

vNM 如何编程?

150 模拟人的认知组织结构,vNM 比 TM 有哪些优势?

"冯·罗依曼瓶颈"指的是什么?

为什么会产生这个问题?

产生式系统(PSs)

PS 的主要构成是什么?

它们是如何组织的?

PS 如何编程?

重写书中产生式示例,避免原示例所存在的错误?

PS 是如何并行加工的,又是如何串行加工的?

PS 与 TMs 和 vNMs 有何区别?

赞成或反对 PSs 的都有哪些观点?

鬼蜮模型

鬼蜮模型的“基本模式”指什么?

鬼蜮模型的一般结构是什么?

鬼蜮模型使用哪两种学习程序?

鬼蜮模型和联结主义系统有哪些相似?

结构类别(I)

定位与分布控制的区别是什么?

间接(相关的)存储与直接(随机的)存储的区别是什么?

存储定位寻址与内容寻址的区别是什么?

哪种机器是串行计算的?

哪种机器是部分串行计算,部分并行计算的?　151

哪种机器几乎是全部并行计算的?

【推荐读物】

概论

对没有专业背景的人来说,关于计算导论性的介绍读物,可见 Glymour(1992)第 12 章和 White(2000)。Boolos and Jeffrey(1989)是较清晰的关于计算的逻辑导论著作。Haugeland(1985)第 4 章是可读性较强的关于机器结构的非技术性导论。

图灵机

关于图灵的生平和职业生涯可见 Hodges(1983)。对图灵机比较全面和完整的介绍可见 Penrose(1989)第 2 章。Barwise and Etchemendy(1999)是关于 TMs 软件及测试的导论。Davis(1958)是最早全面研究图灵机的文献之一,他的选集(1965)包括计算和决策领域的经典著作。对图灵机的经典介绍可见 Minsky(1967),特别是第 6 章和第 7 章。较早关于图灵机和其他系统的比较研究,可见 Gross and Lentin(1970),最新的比较研究见于 Odifreddi(1989)第一部分。在 Dawson(1998)第 2 章中,包含了对图灵机及其与认知科学关系的重要评论。

丘奇-图灵论题

Copeland(1996b)是一篇简明且内容充实的研究文献。1987 年出版的 Notre Dame Journal of Formal Logic 28(4)专门讨论了丘奇-图灵论题。Copeland(1997)讨论了机器的计算功能类别是否等同于图灵机计算的问题(作者认为不是),并叙述了丘奇-图灵论题的普遍(非)公式化问题,以及应用认知科学的非公式化问题。Gandy(1988)详细记录了丘奇-图灵论题的起源和历史发展过程。

冯·诺依曼机

关于冯·诺依曼机的历史著作很多。Augarten(1984)第4章介绍了可储存程序机器的历史,记录了从 ENIAC 到 UNIVAC 的发展过程,还包括有一个关于计算发展重要标志事件的记录。关于 ENIAC 和 IAS 机器的技术性讨论,可见 Metropolis et al.(1980),特别是第四部分。有关冯·诺依曼生平和职业生涯的介绍,可见 Heims(1980)第2章和第14章,Aspray(1990)第1章和第2章。关于寄存机(冯·诺依曼机)理论的简明论述,可见 Minsky(1967)第11章,Clark and Cowell(1976)第1—3章。

152

产生式系统

Minsky(1967)第13章中对 Post 系统有很好的介绍。Newell and Simon (1972,"历史遗补")包含 PSs 的简明历史导论。Barr, Cohen, and Feigenbaum (1981,III,C4)也包含着对 PSs 的讨论。有关 PS 结构应用于多种专家系统的研究见 Kurzweil(1990),有关 PS 结构用于模拟心理功能见该书第8章。

鬼蜮模型

除了文中提到的文献外,关于鬼蜮模型描述及其应用的深入探讨还可参见 Dennett(1991)。

7 表 征

7.1 引 言

为了充分解释人类的认知能力，我们需要把认知看作是包含**表征**“操作”（生成、转换和删除）的过程。那么，认知功能模型必须能够对这些过程进行模拟。这就产生了两个问题： 153

(Q1)

计算模型“操作”（生成、转换和删除）哪些种类的表征？

(Q2)

这些表征如何表达它们的作用——是什么决定着它们所表达的内容？

第一个问题（Q1）被称为**表征种类难题**，第二个问题（Q2）被称为**表征实现难题**。要解答 Q1 需要揭示（i）表征的结构和（ii）计算机表征的主要程式及其特征。解答 Q2 则需要弄明白（i）在怎样的条件下，某物才可称之为一个表征，即如何表征其所表征，以及（ii）究竟什么决定它所表征的内容。

术语小议 提到表征（representations），通常会涉及“符号”、“语义”、“意义”、“指称”、“意向”、“内容”、“关涉”等概念，但是用任何一个词替代都有其优缺点。例如，如果“符号”广义上意谓它所表达的任何事物，那么就可以把表征等同于符号——为了风格多样，本书有时也会以这种方式变换使用。但是在通常情况下，“符号”一词具有非常有限的适用范

154 围，如某人说：我不理解这个句子/岩石上面写着/刻着的一个奇怪符号。“语义”、“意义”、“指称”都可以意指它物，但是它们主要应用于语言系统中的问题。所以，我们采用“表征”一词部分地出于默认——它能够相当广泛地指代各种相关事物，但却较少引入不相关含义。最后，从构词法上看，因为“representation”具有“re-presentations”的结构，所以从中或许可以获得一些真相：我们通常这样解释表征，用某事物 X 意指另一事物 Y，即用 X 重现(re-representing)Y。

在说明这两个问题之前，我们先来分析几个基本问题。首先，尽管我们的兴趣主要集中在“心理”表征上，但是需要注意，事物的表征多种多样，它们的表征模式也相应地多种多样：感知、记忆、意象、自然和人工语言表达(如数字、计算和逻辑符号、笛卡儿坐标)、特殊概念系统表达(如音乐和舞蹈)、布线图、蓝图、地图、绘画、相片、全息图、雕像、仪表和计量器。除此以外，甚至还有临时特设的符号如用缺损的树，或者用一小堆石头做标记。其次，术语不一致会使问题变复杂。我们采用以下常规术语(但是我们需要不那么严格，特别是讨论他人的观点时，因为不是所有人都这样使用)：

原子表征：它们不具有内部表征结构，其中任何一部分都不能表征其他事物。

复杂表征：它们由原子表征构成，组成部分具有表征功能。

组合表征：整体表征取决于其组成成分的表征及其结构关系。

例如，在英语(毕竟是一种表征系统)中，“Venus”这个词指称的是一个行星，但是这个词的任何一部分都不能指称或意谓其他事物(所以它是原子的)。我们也可以使用“the morning star”(或“the last star seen in the morning”)这样的短语来指称金星。它们的各个组成部分都具有独立意义，而整个短语的意义则取决于各个部分的意义以及它们之间的语法关系(所155 以它是复杂的和组合性的)。再来思考一下“kick the bucket(=die)”这个习语。它的组成部分都具有独立意义，但是习语整体的意义并不取决于这些组成部分的意义及其相互之间的语法关系(所以它是复杂的但非组合性的)。要知道这类词组的意义，只需死记硬背就可以了。回到主题，假设某一计算机的寄存器中有以下信息：

1	0	0	1	0	1	1

如果将其看作是字母 k 的 ASCII 代码，那么它就是非组合性的，因为不

在 ASCII 代码表中查询，就不能理解它的意思。如果将这个数串看作是一个二进制的数字(75)，那么它就是组合性的了，其意思可以由二进制位值系统得到。

7.2 表征的多样性：标准的高阶格式

在计算和认知科学中，对于表征的本质，存在着许多不同观点。在认知的数字计算模型中，数据结构在表征中起着主要作用。正如一些程序语言是“高阶”的一样——它们容易使用，可以被翻译(编译、解释)成低阶的机器可以直接使用的代码——一些知识表征系统也是“高阶”的，因为它们简单易用，所以可以将它们翻译成机器可直接使用的基元，如数串、数列、树形图等。典型的高阶表征，包括逻辑(谓词演算)、语义网络和框架。

逻辑：谓词演算(PC)

谓词演算(PC)，又称“量化理论”，在逻辑学中有着辉煌传统，相比其他概念，它的句法、语义和演绎力得到了更透彻的研究。以下的 PC 例子包括了七类表达：谓词、名词、连接词、变量、开句、量词和句子。它们每个都有与之相关的决定其语法性的句法规则以及决定指称和真值的语义结构。 156

谓词

通常将谓词的首字母大写：

F( )：( )是女性

T(,)：( )比( )高

姓名

通常将姓名首字母小写

a：Agnes

b：Betty

句法规则

谓词与特定姓名相结合就成为句子：

F(a)：Agnes 是女性

T(a,b)：Agnes 比 Betty 高

语义规则

一个具有以下形式的句子："谓词＋姓名"，当且仅当姓名所指具有谓词赋予的关系或特点时，为真。例如：

"F(a)"是真，当且仅当 a 具有特点 F 时，也就是说，当且仅当 Agnes 为女性。

"T(a,b)"是真，当且仅当 a 与 b 之间存在关系 T，也就是说，当且仅当 Agnes 比 Betty 高。

连接词

以下空位处是句子可出现的位置

157

__ & __：__与__

__ v __：__：__或__

__→__：如果__那么__

一 __：非__

句法规则

一个连接词与所需句子相结合就构成了复句

F(a) & T(a,b)：Agnes 是女性与 Agnes 比 Betty 高

F(a) v T(a,b)：Agnes 是女性或 Agnes 比 Betty 高

F(a) →T(a,b)：如果 Agnes 是女性，那么 Agnes 比 Betty 高

一F(a)：Agnes 不是女性

一T(a,b)：Agnes 没有 Betty 高

语义规则

一个具有以下形式的句子："句子＋&＋句子"，当且仅当两个句子同时为真，为真("v"，"→"与此类同)。

一个具有以下形式的句子："一句子"，当且仅当句子 S 为假，为真。

术语解注：变量和量词 1879 年，德国数学家、哲学家弗雷格(Gottlob Frege)出版了《概念演算》一书，首次使用变量和量词来分析句子的方法，对逻辑学产生了革命性的影响。令人惊奇的是，这开启了 20 世纪的一个学术探索——逻辑主义者试图证明数学是逻辑的一个分支，是逻辑的结果。1902 年，罗素(Bertrand Russell)发现了弗雷格(完备)系统中存在一个矛盾，并在他与怀海特(Whitehead)的巨著《数学原理》(1910-13)中试图解决这个矛盾。1931 年，年轻的奥地利逻辑学家哥德尔(Kurt Godel)

证明，任何试图将数学纳入逻辑的尝试，如罗素和怀海特的工作一样，必定失败。1936年，图灵使用哥德尔的结果和方法，证明了计算机科学中的一个基本定理：图灵机停机问题的不可解性——一个图灵机不能判断任意输入的另一个图灵机是否停机。弗雷格逻辑，尽管不是特殊的二维计数制，但也是标准的。这里，我们通过量词和变量两个手段来丰富我们的基本逻辑。 158

变量

x，y

句法规则：开句

1. 至少具有一个变量的断言是开句

F(x)：x是女性

T(x,y)：x比y高

T(x,b)：x比Betty高

2. 一个连接词和一定数量的开句相结合仍是开句

F(x) & T(xy)：x是女性与x比y高

T(x,b) → F(x)：如果x比Betty高，那么x女性

量词

以下省略的部分可以由任何合乎句法规则的表达式填充

(Ex) [...x...]：至少存在一个x，x...

(Ey) [...y...]：至少存在一个y，y...

(Ax) [...x...]：对于任一/所有x，x...

(Ay) [...y...]：对于任一/所有y，y...

句法规则：句子 159

开句前加上适当数量（带有适当的变量）的前束量词就是句子：

(Ex) (Ey) [Fx & Txy]：至少存在一个x且至少存在一个y，则x是女性且x比y高

(Ax) [Txb →Fx]：对于任一x，如果x比Betty高，那么x是女性

语义规则：句子

具有以下形式的句子，“存在量词＋开句”，当且仅当存在一个物体具

有开句所意指的特点或关系时，为真。

“(Ex) Fx”为真，当且仅当至少存在一个 X，X 是 F，例如具有女性的特征。

具有以下形式的句子，“普遍量词＋开句”，当且仅当对于任一对象都具有由开句谓词所表达的属性和关系时，为真。

“(Ax) Fx”为真，当且仅当对于任一对象 x，x 是 F，例如具有女性的特征。

“(Ax) [Fx→Txb]”为真，当且仅当对于任一对象，它具有这样的属性，如果是女性，那么则比 Betty 高。

推理规则

&-简化：如果句子“X & Y”为真，那么可推出 X 为真，也可推出 Y 为真(可以推出任何一个项)

取式(Modus Ponens)：如果已知“X→Y”和 X，那么可推出 Y。

这些结构规则表明了表征的形式，并且推理规则使系统演绎成为可能。例如，假设系统需要得到以下表征：

(C)

TALLER(AGNES, BETTY)，就是 Agnes 比 Betty 高

160 而且，以下前提给定时：

(P1)

FEMALE(AGNES) & NOT TALLER(BETTY, AGNES)，就是 Agnes 是女性而且 Betty 没有 Agnes 高。

(P2)

FEMALE(AGNES)→TALLER(AGNES, BETTY)，就是 Agnes 是女性而且 Agnes 比 Betty 高。

系统就可以使用 P1 和 &-简化得到：

(1)

FEMALE(AGNES)，就是 Agnes 是女性。

系统可以使用(1)、P2 和取式得到：

(C)

TALLER(AGNES, BETTY),就是 Agnes 比 Betty 高。

这是一个典型的系统使用 PC 表征系统进行推理的方式。

PC 的优点和缺点

优点

PC 的主要优点在于:(1)具有显性语义;(2)它的形式数学特点已经被研究得较为透彻。另外,(3)它是一个表达思想的自然方式;(4)它是极为“模态”的,因为可以独立于其他语句任意引入或删除任何表述。

缺点

使用 PC 作为认知计算模型的表征法的一个主要缺点是,现实世界中经常同现的信息不能被储存为一个单位。这就给及时有效的激发相关信息带来了困难。一些研究者为此提出了一些能以各种方式有效收集信息的表征方案。其中两种非常普遍:语义网络和框架/脚本。

语义网络(SNs)

语义网络(SNs)起初是作为词义的联想记忆模型而提出来的(所以是“语义的”,见:昆兰(Quinlan), 1966, 1968)。它们大体上是一种图表结构, 161 由结点(圆圈、方块和点)和结点间的连接(弧线、箭头和直线)组成。有的结点表征物体(所谓“个体结点”),有的代表特征(所谓“类属结点”)。连接通常表征了事物之间的联系。语义网络的某一部分可以表征情境。这些都可与 PC 类比:

个体结点::PC 名字
类属结点::PC 一阶谓词
连接::PC 关系谓词
网络片段::PC 断言

图 7.1 表示一个网络样本,此样本——结点和连接组合起来构成了网络片段,网络片段再进行组合就构成了整个网络。

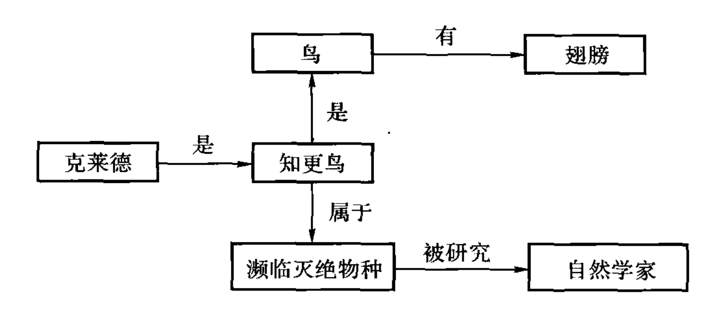


图 7.1 知更鸟网络(来源:Barr, Cohen, and Feigenbaum, 1981: 184)

推理

这里的推理并不是运用推理规则从一个句子推出另一个句子,而是通过网络中的节点进行"激活扩散"。例如,由"知更鸟"开始进行激活扩散,从网络中就能到这样一个基本推理:"知更鸟是鸟"或者"知更鸟有翅膀"。由于"是"(isa)连接的存在,网络中的节点就具有了距其较远节点的特征。例如,在任何一个包括图 7.2 所示的动物网络片段中,低一级的节点可以继承高一级节点的特点。例如,狮子是动物。我们把这种结构恰当地称之为"继承层次"。

162

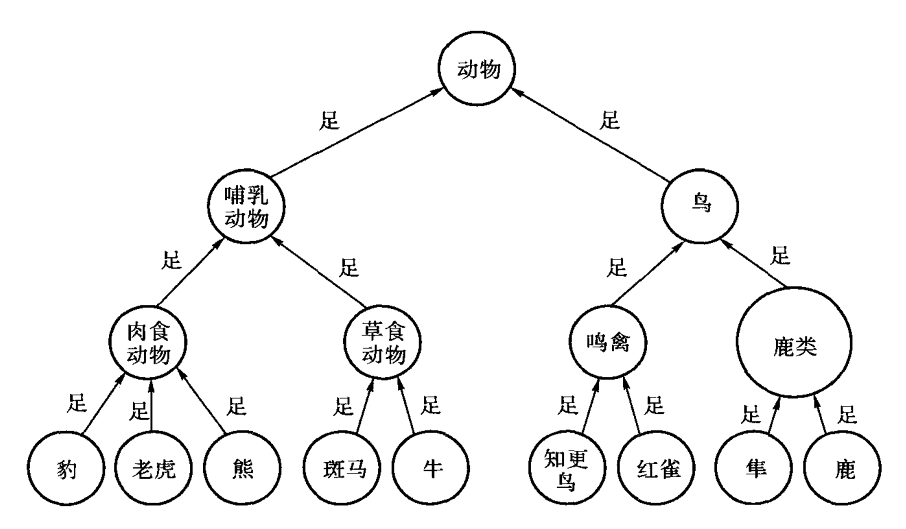


图 7.2 继承层次事例(来源:Staugaard, 1987: 94, 图 3.14)

回答问题

SN 可以回答问题，给它一个部分缺损的语义网络，如果 SN 能在整个语义网络中找到与语义网络片段相匹配的项，那么就找到了缺损部分可能会是什么，也就是**答案**。问题“克莱德有什么?”如图 7.3 所示，SN 从整个语义网络中找到了与这个网络相匹配的部分。如图 7.4，提供了一种答案“克莱德有一个巢穴”。所以，语义网络可以通过激活扩散和匹配加工信息。 163

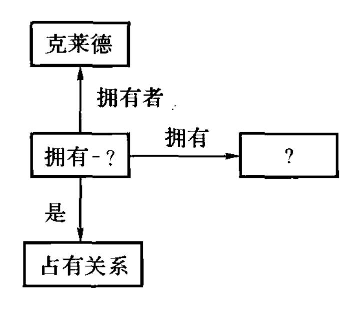


图 7.3　提问网络(来源:Barr, Cohen, and Feigenbaum, 1981: 187)

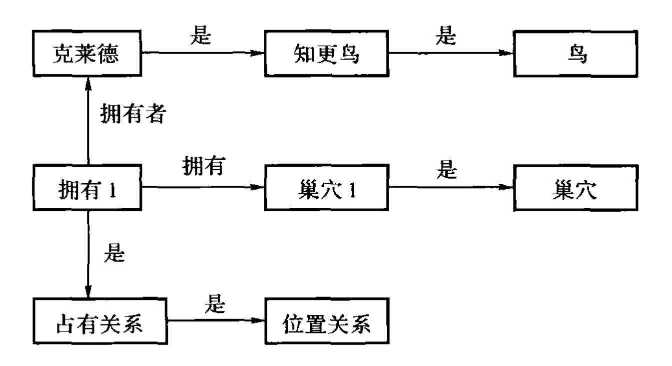


图 7.4　答问网络(来源:Barr, Cohen, and Feigenbaum, 1981: 187)

语义网络的优点和缺点

优点

图 7.1 显示了语义网络的主要优点。在 SN(与 PC 相比较)中，符号“知更鸟”和“鸟”只出现过一次，所有关于知更鸟或鸟的相关信息的节点通过连接建立起联系。这样系统就可以把与知更鸟、鸟等有关的信息集中到一起。

缺点

语义网络的主要缺陷表现在:首先，它们自身带有这样的问题，即不能

准确表达典型性/常态性、析取和否定的概念。如，语义网络如何表达这样的事实：所有的知更鸟都是鸟，但是只有典型的鸟会飞，只有常态的鸟有翅膀，不会飞也没有翅膀的也可能是鸟。语义网络又该怎样表达：克莱德是一只知更鸟或麻雀（不是同时两者）？或者"克莱德不是老虎"？其次，除非受到限制，否则激活会扩散得很远，如：

> 克莱德被自然学家研究（自然学家研究濒临灭绝物种，如知更鸟，但未必研究克莱德）。

所以，获得合适推理内容的途径仍然是必要的。第三，与 PC 不同，SNs 不具有明确的清晰的语义表达。第四，与 PC 不同，我们对 SN 如何进行正确的形式推理尚且了解不多。

框架和脚本

164 明斯基(Minsky)提出的"框架"是最普遍和灵活的数据块结构之一。我们在这里之所以详细引用明斯基关于这个概念的原文，是因为迄今在绝大部分对这个概念的讨论中，均未涉及其中还包含的许多其他思想："这个理论的主旨是：当遇到一个新的情景（或者对当前问题的看法发生了实质性改变）时，人们会从记忆里选取一种结构叫做**框架**。这个储存在记忆中的框架可以通过变更细节来适应现实情况。框架是一种表征定势化情境的一种数据结构，如处于某个客厅中或参加一个孩子的生日聚会。在框架中，有些信息是关于如何使用这个框架的，有些信息是关于可能的继发事件的判断，还有一些信息是关于如果这些判断没有被确认应该如何做。我们可以把框架看作是一个由结点和连接构成的网络。框架的高层是固定的，表征在某一情境中通常发生的事件。框架的低层有许多**终端**——需要被具体例子或数据填充的插槽，每一个终端对它的赋值都有条件（这些赋值自身通常是较小的子框架）。简单条件由**记号**标明，这些记号表明终端的赋值可以是人、满足某种条件的物，或指向某一类型子框架的标记。较复杂的条件可以用多个终端的赋值之间的关系来表达。多个相关框架的集合构成了框架系统，在系统中框架的转换反映了发生重大事件的影响……**系统中的不同框架可以具有同样的终端**，这一点非常重要，它使系统能够协调从不同角度收集到的信息。理论的现象学力量产生于期望的内容以及其他各种推测。**框架的终端通常由'缺省'赋值填充**，所以框架可能包含很多推测的细节，这些细节并没有得到具体情境的证实。框架系统反过来通过信息检索网络联系在一起。当一个框架不能适合实际情境时——也就是不存在符

合终端条件的终端时——它就会提供出一个替代框架……当某一框架表征某一情境时，**匹配**过程就试图为每个终端赋值，与每个位置上的记号相一致。这个匹配过程一方面受到框架相关信息（包括如何处理意外事件的信息）的制约，另一方面它也受到系统对当前目标的认识的制约”（Haugeland，1997：111-37）。

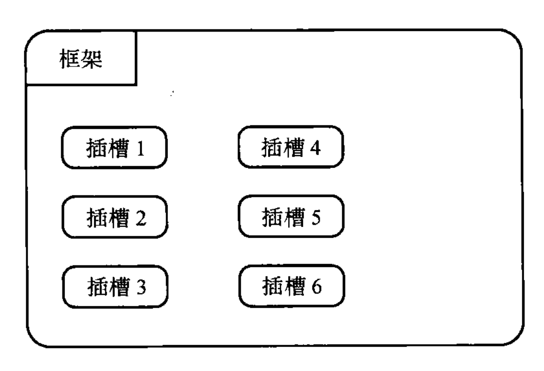


图 7.5　一般框架结构（来源：Staugaard，1987：97，图 3.16）

下面我们对这段较长引文中的一些主要观点进行解释。框架表征了关于某一（某类）对象、特征以及情境的**模块化**信息（明斯基将框架概念拓展为事件序列，但我们通常将事件序列称作“脚本”，在这里我们也将其称之为脚本）。框架将现实世界中的相关信息通过框架中的插槽“整合”到一起（如图 7.5）。 165

插槽可以由经验、缺省值或指向其他框架（和脚本）的标记填充。下面通过几个具体例子来说明这些观点。

房间框架

例如，我们回想对于房间的认识。首先，我们知道房间会有一种通常结构，在框架理论中，这意味着房间框架中有许多插槽标记一个房间通常具有的组成部分（如图 7.6）。其次，我们知道还有很多种不同类型的房间，所以存在另一套指向这些典型房间的标记（如图 7.7）。

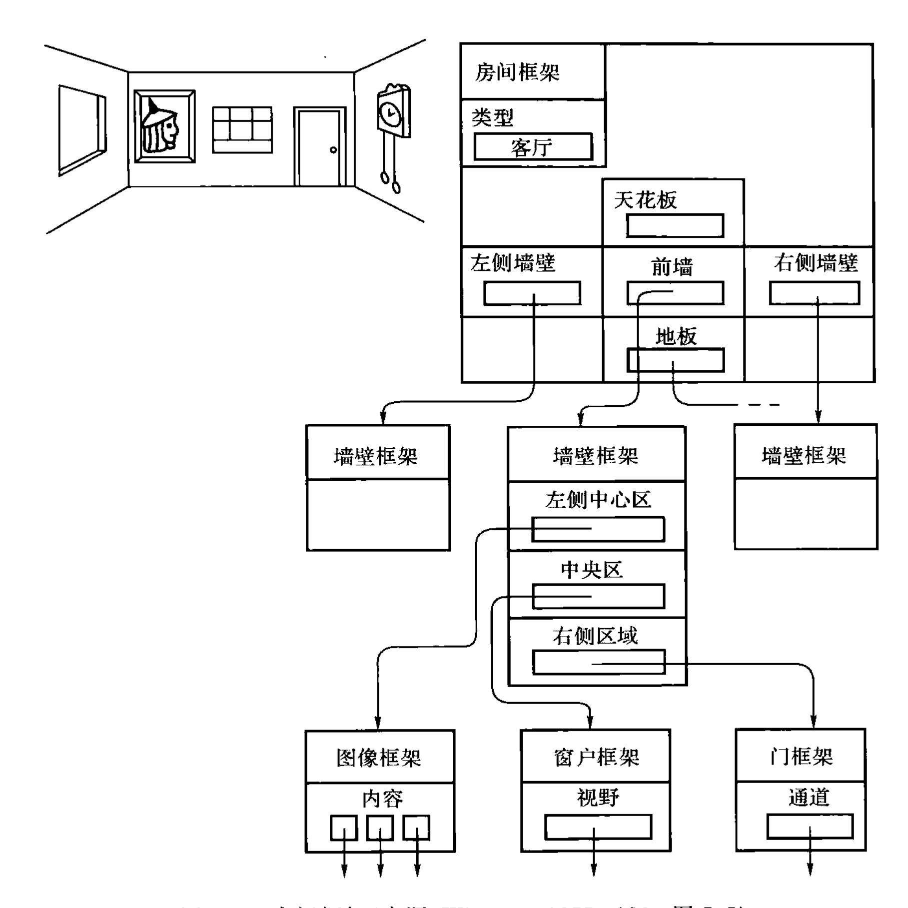


图 7.6　房间框架(来源:Winston, 1977: 186, 图 7.5)

餐馆框架

再来考虑我们对餐馆的一般认识。首先我们知道餐馆是一个特殊的用餐场所。在餐馆框架中,有很多填充具体信息的插槽。例如(见图 7.8),在这样一个餐馆框架里,用餐方式和地点等插槽需要通过其他框架填充。填充框架的信息越详细,机器就"理解"得越充分。

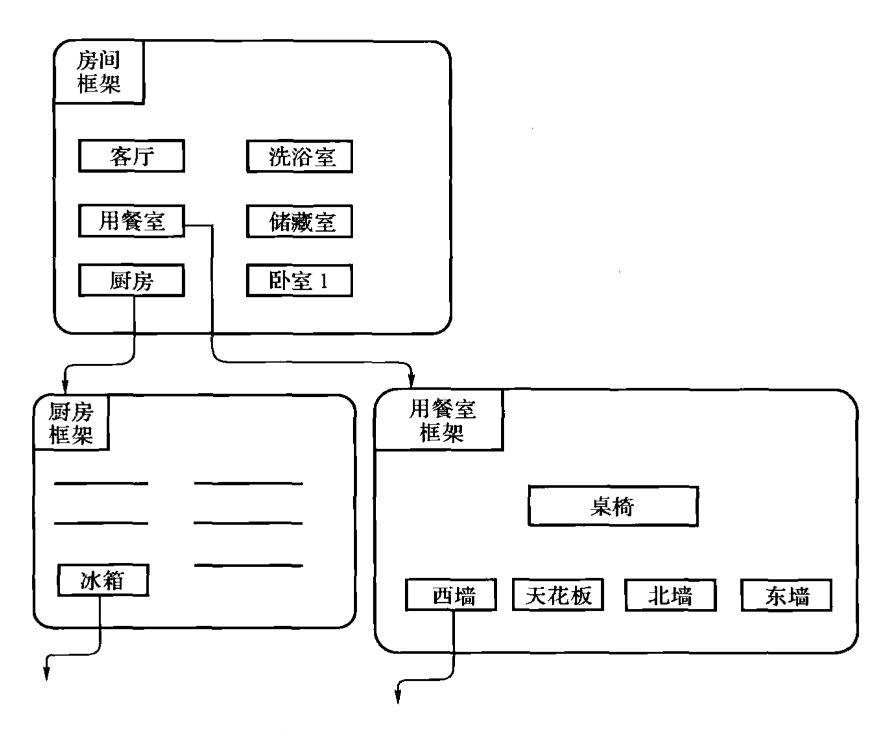


图 7.7　房间类型框架(改自:Staugaard, 1987: 100-1, 图 3.19、图 3.21)

脚本/行动/框架

脚本(也叫做行动框架)是关于原型事件的序列,或者记录了关于事件序列的一般性**预期**,或者是形成指导行动**计划**的基础。它们可以被框架调用,或者用来解决某种问题,或者用来实现某种目标。我们将阐述脚本的两种用法:一种是脚本可由框架调用,以形成关于某种事件序列预期的基础;另一种是脚本可作为解决某种问题或者实现某种目标的行动计划基础。

在餐馆用餐脚本

对于脚本的第一种用法,我们再回到刚讨论过的餐馆框架。餐馆框架中的"事件序列"插槽调用了"在餐馆用餐"的脚本。这个过程是怎样的呢?如图 7.9 所示。 167

这个脚本包含了两类信息。第一,脚本的组成"成分":道具、角色、视角等。这些成分表明了脚本中的参与者或"行动者",以及他们活动的场景。第二,事件的一般时间序列。图中脚本记录了这样的信息,例如,在正式用餐的餐馆中可以在用餐完毕后买单,但是在快餐店则需要在用餐前付账。

168

一般餐馆框架
经营类型:商业服务
用餐方式:
　范围:　　　(自助、自请入座、静候入席)
　缺省:　　　静候入席
　如果一需要:　**如果**餐馆里有点餐后即刻付账的柜台,**那么**用餐方式为快餐。
　　　　　　　如果需要用堆栈起来的餐盘取餐,**那么**用餐方式为自助餐(Cafeteria)。
　　　　　　　如果需要等待侍者示意,或者已有预订,**那么**静候入席。
　　　　　　　除上述情况外,自请入座。

地点:
　范围:某地址
　如果一需要:(看菜单)

名称:
　如果一需要:(看菜单)

主食类型:
　范围:(汉堡包、中国风味、美国风味、海鲜类、法国风味)
　缺省:美国风味
　如果一需要:(餐馆里更新的可供选择的主食)

经营时间:
　范围:每天经营的时间段
　缺省:除周一外每天晚上开放

付款方式:
　范围:(现金、信用卡、支票、按劳相抵)

事件序列:
　缺省:在餐馆用餐脚本

备选方案:
　范围:拥有相同主食风格的所有餐馆
　如果一需要:(寻找拥有相同主食风格的所有餐馆)

图 7.8　餐馆框架(来源:Barr, Cohen, and Feigenbaum, 1981: 217-18)

餐馆用餐脚本
　道具:　　　(餐馆、饭钱、食物、菜单、餐桌、餐椅)
　人物:　　　(顾客、侍者、出纳员)
　视角:　　　顾客
　事件时间:　(餐馆的营业时间)
　事件地点:　(餐馆所在地点)
　事件序列
　　首先:　进入参观脚本
　　然后:　如果(静候入座 或者 餐位预订)
　　那么获得侍者注意脚本
　　然后:　入座脚本
　　然后:　点餐脚本
　　然后:　用餐脚本,除非(因上餐迟延)不满而离开餐馆脚本
　　然后:　如果(菜肴非常美味)
　　　　　　那么称赞厨师脚本
　　然后:　付款脚本
　　最后:　离开餐馆脚本

图 7.9　餐馆用餐脚本(来源:Barr, Cohen, and Feigenbaum, 1981: 218-19)

移动圆盘脚本

脚本的第二种用法是用于解决某个问题或实现某种目标。这样一个脚本(或行为框架)可以具有如图 7.10 所示的简化形式。

169

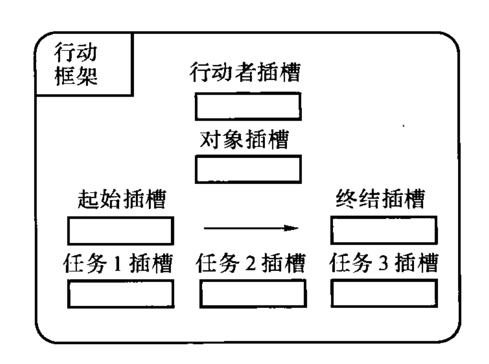


图 7.10　一般脚本或行动框架(改自:Staugaard, 1987: 103,图 3.23)

我们想象一个"**行动者**"针对某**对象**实施了某种行为,即完成了使对象从起始状态发展到终结状态的任务。例如,想象一个机器人把一根柱子上的圆盘移动到另一根柱子上,并且保持圆盘的序列不变(如图 7.11)。机器人按照脚本从顶部圆盘 A 开始,将其移动到中间的柱子上,这个活动的脚本/行动框架如图 7.12 所示。

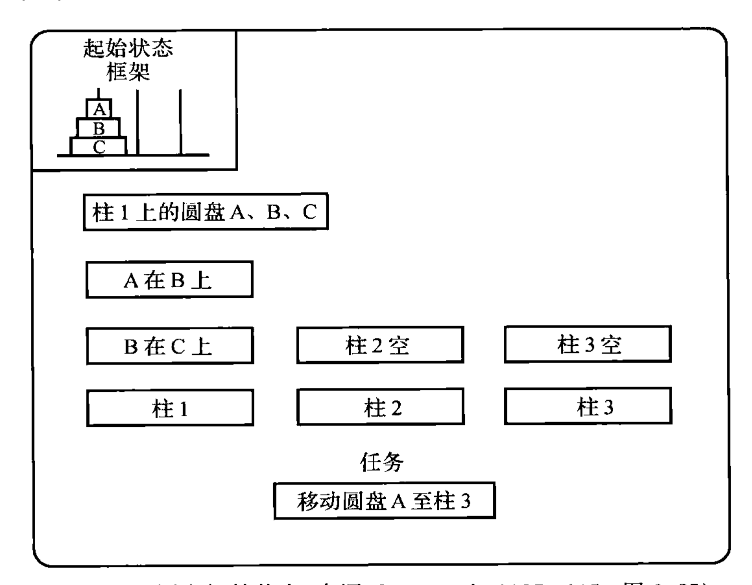


图 7.11　圆盘初始状态(来源:Staugaard, 1987: 105, 图 3.25)

170

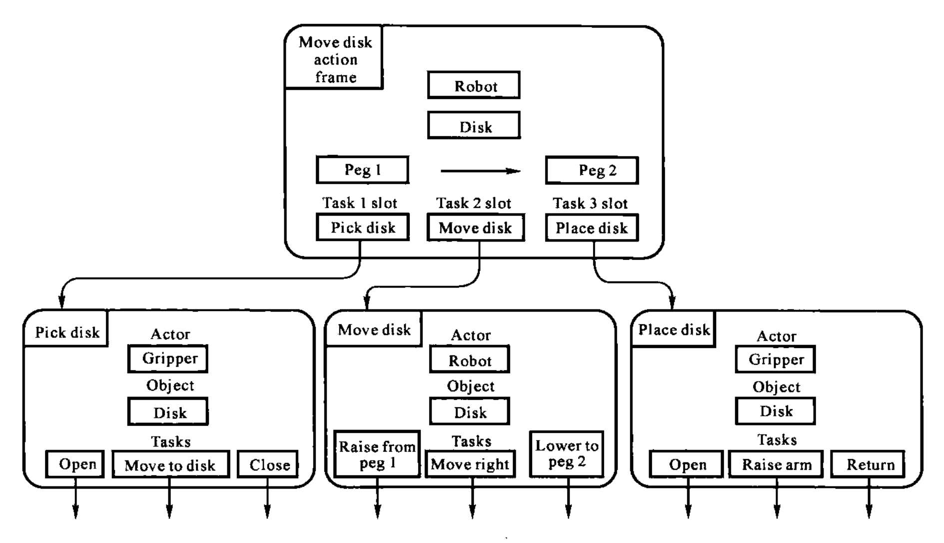


图 7.12　一般脚本或行动框架(来源:Staugaard, 1987:104,图 3.24)

更详细的脚本说明接下来解决这个问题的动作:继续将剩下的圆盘(从上自下)移动到中间的柱子上,这样就形成了 C, B, A 的一摞。然后将这些圆盘(自上而下)从第二根柱子移动到第三根,这样就形成了 A, B, C 的一摞,问题得以解决。尽管这是个很简单的人工智能实例,但是它却说明了脚本是如何用来指导事件序列的,而不仅仅是记录事件序列的预期。

框架/脚本的优点和缺点

优点

框架/脚本的主要优点在于可以将相关信息容易地组合在一起,所以,171 框架、脚本、图式等在人工智能甚至认知心理学领域都起着重要作用。

缺点

框架脚本的主要缺点是:(1)它们(与语义网络一样)不具有明显的语义特征;(2)它们(与语义网络一样)没有关于范围和界限的一般性理论;(3)有些信息不属于任何框架或脚本,而是我们普通常识的一部分,如,买东西或得到服务需要付钱,或者放开无支撑的圆盘时它会下落。

高阶方案:总体评价

这三种表征程式各有优点与缺点,如图 7.13 所示。我们所需要找到的

是具备所有这些优点而没有这些缺点的表征体系。

	PC	SN	F/S
明确表达语义	是	否	否
一般性理论	是	否	否
有效组织信息	否	是	是

图 7.13 高阶表征的优缺点

7.3 数字计算表征的本质

本章开头提出的第二个问题涉及表征的本质。在数字计算理论看来，**表征如何表达**？或者说，**是什么决定着表征所表达的内容**？通过以上对三种表征形式的分析，我们可以提出这样的问题："T(a,b)"(＝Taller (Agnes, Betty))怎样表征了 Agnes 比 Betty 高这样一个事实？语义网络左边的部分 172 怎样表征了"克莱德是知更鸟"(Clyde is a robin)？房间框架怎样表征了房间(而不是汽车)？餐馆脚本是如何表征在餐馆里用餐(而不是看牙医)的一般事件序列的？然而，令人惊异的是，这些符号系统的表征并不具有清晰的组合性质，我们也可以说，它们没有**明确的**组合语义特征。对于我们讨论过的所有系统来说，它们所表征的内容并不属于计算理论。我们之所以知道系统的**意向**——程序员对它的指令，是因为我们看到了显示在系统上的语句，如英文。这些语句帮助了我们，但是解读这些语句却不是**计算机**程序的一部分。我们也可以这样问：这些表征对于**计算机**而言，而不是**我们**，说明了什么或者意味着什么呢？

解释语义(IS)

一个可能的假设(Cummins, 1989, 第 8 章和第 10 章)是，**数字计算表征因为与它们所表征的事物具有同构性**(isomorphic)，从而具有了表征能力。根据康明斯的说法，计算机中的这种表征类型非常特殊，应该给予一个特别的名称，即"模拟表征(s-representation)"。

实例

设想，我们现在想解释袖珍计算器是如何进行加法计算的。加法就是正确地使用(＋)号进行计算，那么袖珍计算器又是如何做到的呢？通过操作表征数字的数码。那么，我们是怎样使用计算器做加法计算的呢？是通过这样的按键顺序：先“归零”，第一“数项”，再按“加号”键，第二“数项”，之后是“等号”键，结果就会显示出来。要想使一个机器可以做加法计算，我们只需设计好让机器能够表征加法操作和两个数字：

1. 不同的按键是不同数字的表征。

2. 屏幕显示的内容也是数字的表征。

3. 计算器装置使屏幕显示按键操作的结果。

4. 如果按了表征 m 和 n 相加记号的按键，那么屏幕就会显示 $n+m$ 的记号，然后得出结果。(Cummins，1989：90)

表征

173 但是按键、屏幕显示和数字之间的关系，倘若不是因为人的解读，那么又是从何而来？康明斯认为，以上的操作系列与加法(＋)是**同构性的**，所以才能够成为数字的表征。如果两个按键被误标以及/或者这两个按键连接屏幕的显示线交错了，我们就必须弄清楚正确的符号表征，即分析哪种符号一数字的关联能够满足加法运算。如康明斯所说：“在某种意义上……它(加法计算器)能够表征数字，是因为它能够进行加法计算：我们认为它能够表征数字，只是因为它在某种解释下模拟了‘＋’号”(1989:93)。正是在这种意义上，图表或方程式能够表征数据，或者抛物线表征了抛体的轨迹。

问题

但是因为各种不同事物都可能与同一种表征具有同构性，所以这种假设就面临这样一个问题，即怎样使表征能够正确地表征它所要表征的事物。若一台机器可以被解释为是做乘法运算的机器，那么它也可以被解释为在做加法运算。的确，如果同构条件充分，一种表征可以表征的事物的数量确实是无限的。“目标选择”可以作为解决这个问题的一个方法。此方法包括两个步骤，首先我们需要知道所有的同构解释，然后再根据计算目的从中选取“正确”的解释。显然，第二个步骤并不是计算的，是我们需要引入计算领域之外的东西。因此，我们又回到了这个计算表征难题(Q2)——表征的本质是什么：表征是如何具有表征功能的？——仍然无解。

逻辑关系难题

我们在前面(第4章)曾提出了简单探测器语义的三个难题:因果(或纵向)难题、质性(或横向)难题、错误表征(或析取)难题。我们已经回顾了一些标准高阶计算表征系统,所以可以提出第四个难题——**如何解释表征之间的逻辑关系**(讨论“蛙”模型时,我们也可以提出类似的问题,但是对于蛙而言,似乎讨论表征间的逻辑关系是没有意义的)。对于DCTM而言,我们找到了它成立的更为充分的理由,因为逻辑关系已经得到深入研究,并在计算机科学中得到充分实践——显现了谓词计算(PC)作为知识表征图式的优势。 174

但这里还存在一个重要问题。在讨论PC时我们知道,系统由两套规则决定:确立和推导句间关系的句法规则,分配指称和赋予真值的语义规则。有的逻辑关系,如推论和论证,是“句法”的。有的则是语义的,如真值和蕴涵(P蕴涵Q,指当P为真时,Q必为真)。PC中的真值取决于指称,由于这个概念将表征与现实世界联系在一起,所以就不受计算解释的限制了。但是,自从20世纪30年代早期开始,对于很多表征系统,若系统中的表征P蕴涵另一个表征Q,那么系统中就必然存在一个从P到Q的论证(由规则得到的推理)。也就是说,系统有某种“完备性”。当一个系统具有这种完备性时,演绎理论(证明理论)就可以替代真值关系理论(模型理论)——内部论证关系**模拟**外界真值的关系。数字计算机擅长的是确定表征间推理关系的轨迹,这就增加了计算表征不仅具有句法逻辑关系,而且还具有语义逻辑关系的可能。例如,在下一章中我们会提到一些学者,他们认为计算机的这种通过内部推理关系模拟外界真值关系的能力,使我们有理由认为计算机能够表征某种语义内容。

【思考题】

引言

关于表征的两个主要问题是什么?

原子表征、复杂表征和组合表征分别是什么?举例说明。

逻辑:谓词演算(PC)

用PC表达下列陈述:

Betty是女性

Betty比Agnes高

有些人是女人

任何事物都比 Agnes 高或 Agnes 不是女的。

PC 表征系统所具有的主要难题是什么？

语义网络(SN)

175 语义网络是怎样解决谓词演算的主要难题的？

用语义网络表达下列句子：

Agnes 比 Betty 高与 Betty 是女性

语义网络是通过什么机制进行推理的？

语义网络是通过什么机制回答问题的？

语义网络存在什么难题？

框架(和脚本)

什么是框架？

什么是脚本？

脚本的两种主要用法是什么？

给出一个“汽车”的框架。

给出一个“看医生”的框架。

框架和脚本的优点是什么？

框架和脚本的缺点是什么？

表征的本质

对于如何表达表征的作用问题，三种表征系统存在什么共同难题？

【推荐读物】

引言

Cummins (1989)第 1 章对表征所涉及的问题作了很好的介绍。Cummins (1986)讨论了信息如何“存在于”系统中，但并未被“明确表征”。要了解数据结构方面的知识，可以参考 Tremblay and Sorensen (1984)。

数字计算表征的多样性

Barr, Cohen, and Feigenbaum (1981)回顾了本章所讨论的几个表征方案，也涉及了一些我们没有讨论到的表征方案。Rich (1984，第 7 章)论述了语义网络、框架和脚本。Cherniak and McDermott (1985)的第 1 章也讨论了表征及其在人工176 智能(AI)领域中的作用。Staugaard (1987)第 3 章从机器/AI 的角度，对我们所有讨论过的表征系统作了很好的回顾。Partridge (1996)讨论了本章涉及到的表征方案，以及 AI 中知识表征的一般问题。Thagard (1996)的第 1 章将表征与计算联系起来，然后(第 2 章到第 4 章)从认知科学的角度探讨了谓词逻辑和框架。

逻辑:谓词演算

Jeffrey (1991)是一本简明的逻辑学导论。Rich (1983)的第五章讨论了谓词演算的表征。Mylopoulos and Levesque (1984)特别讨论了一些与 PC 相关的内容。McDcrmott (1986)和 Hewitt (1990)批评性地分析了 AI 中的逻辑作用。

语义网络

Quinlan(1966)是探讨语义网络,在记忆模型中的心理作用的早期著作,其中部分思想也见于 Quinlan (1968)。Woods (1975)详细论述了语义网络的理论合理性。Brachman (1979)讨论了知识表达层次的问题,也包括了一些历史性回顾,以及对其不足提出的建议。对于语义网络在心理学上的应用,可参考 Norman and Rumelhart (1975)。

框架/脚本

Minsky (1975)是框架研究的经典之作,直至现在也是推荐书目之一。Winograd (1975)和 Hayes (1980)结合机器表征的一般问题讨论了框架。Schank and Abelson (1977), Eysenck and Keane (1990)提供了研究框架和脚本的心理学方法。

数字计算表征的本质

McDermott (1976)和 Hayes(1979)对表征中的标记问题方面给予了特别关注。McClamrock (1995)尝试将计算特征融入现实世界。

解释语义

Haugeland (1985)的第 3 章介绍了解释语义学,Cummins (1989)和 Cummins (1996)分别作了详细的探讨和修订。Horgan (1994)作了批评性分析。

逻辑关系的难题

177

对于外界真值关系如何从内部证明关系获得的问题,Tarski (1969)是一部可读性较强的权威著作。Rey (1997)第 8.2 节从认知科学的角度简要讨论了计算中的演绎关系。很多优秀的关于符号逻辑的教材,都讨论了逻辑系统中完备性和不完备性的问题,其中 Jeffrey (1991)是一部颇富洞见的著作。

8 心智的数字计算理论

8.1 引　言

178　认知计算模型所蕴含的主导思想是：**认知主要涉及对我们周围世界的心理表征进行心理操作（生成、转换和删除）的概念，而计算机就是这样一种能够自动操作符号的机器**。我们辨别了关于数字计算机的两个主要问题：结构和表征。我们把结构分解成两个主要部分：记忆和控制，并根据这两个概念又区分了一些重要结构。我们还考察了一些在数字人工智能中有影响的表征图式。现在回到认知方面。我们如此认真地对待计算机这个"类比"，还有其他更多的具体原因吗？也就是说，从字面看，为什么**认知确实是计算的一个种类**？心理表征的实质是什么？计算是如何涉及意识的？**认知**结构的种类有哪些？

这些问题即便未必需要回答，但至少也需要面对它们。目前，认知心理学和认知神经科学正致力于研究这些问题。这里，我们不详细讨论细节问题。我们感兴趣的是那些使"信息加工"心理学得以成立的一般计算框架。既然存在多种可能的计算机结构，那么，那种认为计算机模型只具有某一特定认知结构的观点就是错误的。在接下来的几节中，我们将忽略不同种类计算机在结构细节上的差别，而关注它们都是根据一般计算原则对表征进行自动操作的事实。在本章的结尾，我们将回到对认知结构问题的讨论。

8.2 从心智的表征理论到心智的计算理论

DCTM 不是**凭空**出现的,我们最好把它理解为更为一般的理论的一个特例,即所谓心智的表征理论(RTM)。 179

心智的表征理论

我们首先把心智的表征理论基本形式作如下阐述:

(RTM)

1. 认知状态是具有内容的(心理)表征的关系。
2. 认知过程是对这些表征进行的(心理)操作。

根据这种阐述,"相信"是一种表征的关系,"计划"也是一种,"意愿"又是另一种表征关系,因为不同的表征构成了不同相信的、计划的和意愿的事物。有两种具有影响力的哲学传统在 RTM 这里交汇在一起——一种来自于休谟,另一种来自弗雷格和罗素。每个传统都具有它的优点和不足。

休 谟

休谟认为,心理表征就是当我们在处于诸如相信、意愿和计划等心理状态时所"持有"的"观念"。在休谟看来,心理过程就是观念联结的序列,人们就是通过援引这些相关观念来解释行为的。然而,这并没有说明联结观念的"语义"是什么——它们是"指向什么"的,以及它们是"如何指向"它们所"指向"的?换句话说,休谟对观念和心理过程的解释并没有告诉我们,例如,一个信念何以为真,以及信念中的观念何以确定对某对象的指称和指向。休谟似乎认为观念就像**图像**一样,指向与它们**相似**或者最相似的东西。但是,相似性作为一种表征理论仍存在各种各样的难题。首先,把观念用图像来类比,可能存在概括不足或者过于概括的问题;第二,没有真正解释心理指称的问题;第三,不能判断事物的真假。对这些问题作一些详细的考察是有意义的。对于**第一个问题**,我们知道人们的很多观念并不与某物的图像相似(例如,像"正义"这样抽象的概念),它们更像是**概念**。就算当它们是图像的时候(当想象埃菲尔铁塔时,会具有埃菲尔铁塔的一个形象),有时也会因为太过于具体而不准确。对于埃菲尔铁塔的图像,一个人可能把它表 180

示为一些大概的形状、方位、相对大小以及观看的角度(从旁边看,不是从上往下看)等,但是这些特征没有一个是我们能够想起埃菲尔铁塔时必不可少的。在另一方面,有些图像是相当难以确定的(也有人说“模棱两可”)。维特根斯坦(Wittgenstein)的一个著名例子是拨火棍图像,设想拨火棍一端放置在另一端的前面并形成了一个斜面,那么,这根拨火棍是要站立,还是要倒下?这两种可能性在这幅图像中都是存在的。对于第二个问题,一般来说,相似性并不是表征的充分条件。你的左手与它自身相似(与你的右手也相似),但是其中并没有表征关系。一个政治卡通人物可能与尼克松的模仿者而不是他本人更加相似,但是我们不会说,这个卡通人物因此表征的是这个模仿者,而不是尼克松本人。最后,对于第三个问题,图像本身不能判断真假——真假问题主要针对的是“句子”和“命题”,而不是“名词”。例如,我们不能说,“埃菲尔铁塔”是真还是假,但我们可以说,“埃菲尔铁塔是由古斯塔·埃菲尔建造的”是真还是假。

弗雷格和罗素

弗雷格-罗素的理论试图解决对休谟的这些批评,至少包括了上述的第二和第三个问题。罗素(Russell, 1918)把“命题态度”引入了心智理论,即任何心理状态都涉及一种对命题(P)的态度。典型的例子包括:相信P,意愿P,计划P。弗雷格(Frege, 1892, 1918)认为,相信P包括两个部分,首先,有一个被相信的命题P,它提供了心理的内容——指向什么。这个命题可以是真也可以是假,所以相信的内容可以是真也可以是假,因而相信本身就可以是真也可以是假。其次,在人和命题之间存在一种弗雷格称之为“领悟”的关系。对一个命题不同的“把握(grasping)”方式会产生不同的态度,例如,命题P=火星上有生命存在,一个人可以相信P(火星上有生命存在)、希望P、惧怕P、意愿P等。弗雷格从来也没有说过“把握”一个命题到底是什么意思,虽然他也曾经将“把握”一个命题比拟为感受一种情境,但最后他宣称它是一个“谜”——如果考虑到弗雷格认为命题是一种抽象存在,那么谜的说法就使问题更加突出了。设想一个具有时空维度的实体,例如人,如何能够“把握”一种抽象存在呢?即便我们能够解释对这个命题的把握,但又如何能够解释是对这个命题的把握,而不是对另外一个命题的把握导致了我们的行为呢?如果我渴了并且相信冰箱里有啤酒,那么,这个信念对我产生行为具有因果关系,而且用这个信念能解释我的行为。对一种抽象存在的把握,如何能够起到这样的作用呢?这里主要考虑的不是抽象命题本身,而

181

是抽象命题与心理的关系，而它能够起到心理解释的作用。

命题态度

在本章后面和下一章，我们将会看到“命题态度”对于 DCTM 的重要性。现在，我们只作简要讨论。命题态度（相信、意愿、计划等）同时具有一些共同特征，也有一些区别。这种思路对我们的讨论多有助益。

共同特征

(1)如前所述，弗雷格和罗素认为，认知状态之所以称为“命题态度”，部分原因是它们可以被分解为两个部分：命题内容以及对这个命题内容的“态度倾向”（也就是塞尔称之为“心理模式”的部分（Searle，1979，1983））。同样地，命题态度的特征一般来说具有以下形式：“xAp”，这里“x”代表人，“A”代表一种态度（相信、意愿、计划），而“p”代表命题内容。我们通常用表示态度的**动词**加上表达命题内容的**句子**来表示一种命题态度：

态度动词 + *命题句子*
相信(believe) 火星上有生命((that) there is life on Mars)

有时候我们需要对语言做适当变化以与某些态度动词相匹配，但我们仍然能够从上面这些例子看出对每个命题的态度来。

态度动词 + *命题句子*
渴望(desire) 火星上应当有生命((that) there be life on Mars)
希望(want) 火星上会有生命(there to be life on Mars)
计划(intend) 我要去火星(that I go to Mars)

差异特征

(2)不同的态度倾向之间有一个很重要的不同，即所谓的“匹配的责任”(Austin)或“适切的方向”(Anscombe，Searle)。虽然当命题态度被满足时，它们都与现实世界相适切，但适切方向却因态度倾向不同而不同。例如，信念表达的是人的大脑中已有的对世界图景的认识，如果它与现实世界相匹配，那么就为真；反之，信念就为假。我们会说信念（以及与信念类似的其他态度倾向）具有一种从**心理－到－世界**的适切方向。与之相对，意愿和计划是我们给未来世界勾勒的蓝图——如果现实世界的发展确实与其相符，未来世界的图景与我们所意愿和计划的样子一致，它们就被满足。我们说意愿和计划具有一种从**世界－到－心理**的适切方向。一种态度倾向总有并且 182

只有一种适切方向，也就是说，不可能出现某些信念具有一种适切方向，而另一些信念却具有不同的适切方向。这是因为命题态度就决定了态度倾向会总有一种适切方向。还因为命题内容 P 表达了世界是或将是的样子，也就是**命题已经决定了适切条件**（两种方向中的任何一种）（见图 8.1）。

适切方向：	心理一到一世界	世界一到一心理
命题态度：	×{相信, 知道}p	×{渴望, 打算}p

图 8.1　命题态度和适切方向

（3）不同态度倾向之间另外一个潜在的差异与态度倾向的体验特征有关，或者说是否具有那种独特的体验特征。例如，如果你**害怕**房间里有一个闯入者，你可能会有一种特别的感觉，这种感觉绝对不会与**渴望**有一个闯入者的感觉相混淆。但如果你**相信**火星上有生命（或者同样的事情，即相信房间里有一个闯入者），是否也伴随一种特别的“相信的感觉”，如同在害怕与渴望时会伴随产生某种特别的感觉一样？更进一步说，这些“感觉”是态度倾向的必要条件，或者只是一种伴随物？这些都是目前认知科学争论的话题，我们将在后面回到这些话题的某些方面。

评论

休谟的理论（观念和相似性的联结）在提供心理解释方面具有优势，但它在描述表征的内容方面却没有可行性——（1）图像过于具体，以及只能局限于某一类心理表征，因而不足以用来作为所有心理指称的模型；（2）图像也不能说明命题态度的真假、满足或不满足等问题。弗雷格和罗素的理论解决了一般意义上命题内容的真假等问题，但却没有提供心理解释。我们需要一个既能提供心理解释又能解决表征指向内容及真假问题的理论。

183

心智的数字计算理论基础

在我们详细阐述和论证完整的心智的数字计算理论（DCTM）之前，我们首先简要介绍关于这种计算进路的一些历史背景和影响因素，然后提出 DCTM 的基础形式。

历史背景

1947 年，图灵（Alan Turing）作了一个题为“智能机器”的讲演，他提议要“研究机器能否具有智能行为”。接着他设想了各种“制造‘智能机器’”的

方法(原文带引号)。特别值得一提的是,他还研究了训练机器的方法,正如他笑言:"期望一台从工厂直接出来的机器和一名大学毕业生进行平等竞争是不公平的。"

"图灵测试"

图灵在其1950年的著名论文《计算机器与智能》中介绍了一种"模仿游戏",也就是人们熟知的判断机器是否具有智能行为的"图灵测试"的简化形式。下面是图灵本人(Turing, 1950)对这个问题的出发点、测试设计,以及对测试的断言:

> ***出发点*** 考虑这样一个问题:"机器能够思维吗?"……这个问题可以用另外一个问题来代替,用作替代的问题与它紧密相关,并且是用没有歧义的词语来表达的。
>
> ***模仿游戏*** 这个问题的新形式可以用所谓的"模仿游戏"来描述。有三个人玩这个游戏:一个男人(A),一个女人(B)和一个提问者(C),提问者性别不限。提问者在另外的房间,和其他两人分开。游戏的目的是让提问者来分辨哪一个是男人,哪一个是女人。提问者以标签X和Y来代表他们,在游戏的最后,提问者说出"X是A,Y是B"或者"X是B,Y是A"……理想的方案是让一个电传打字员与两个房间进行联系……现在我们问:"如果一台机器代替A来玩这个游戏,结果会发生什么事情?"在这种情况下,提问者在作判断时发生的错误率,是否也像没有用机器替代前一样呢?这些问题就代替了我们最初的"机器能够思维吗?"这样的问题。 184
>
> 新的问题对人的身体特征和思维进行了分离,使得C只能通过对话识别是人还是机器。
>
> ***赢得游戏*** 我相信在未来大约50年内将有可能会出现这样的编程机器,它的储存容量大约是10^9比特。这种机器在参加模仿游戏时,会表现得非常好,一般的提问者在提问5分钟后进行辨认识别的正确率不超过70%。我相信最初的问题"机器能够思维吗?"会变得毫无意义而不值得讨论。不过我相信,到本世纪末,这些术语的使用和教育的普及将会得到很大的改观,人们能够自由谈论机器思考的问题而不会招致非议。

虽然图灵讨论了使一台数字计算机能够思维的条件,但这本身并不意味着我们可以宣称**人类的思维**就是计算。也就是说,某些数字计算是"思维"(图灵经常在这里使用引号),人的认知是思维,但人的认知却可能不是

数字计算。我们也很难找到图灵有过人的思维是某一种数字计算这样的断言。在他 1947 年的讲演中(前文已提及),图灵有过这样的鼓动性言论:“所有的这些表明,婴儿的脑皮层是一台未加组织的机器,它能够经由适当的干预训练而得到组织。组织活动可能导致机器的调试而变成普适机器或者类似的东西”(Turing, 1947:120)。这个观点似乎意味着,脑皮层开始时处于一种由相互联结的神经元所组成的混沌状态(一种联结主义机器?),通过训练/教育,它变得有组织,接近于一台普适图灵机。由此可见,是由于脑皮层具有计算状态,所以使大脑具有了认知状态,这样我们就获得了心智计算理论的一个早期版本。

在人工智能以及部分认知科学领域,“图灵测试”被当作是对思维(智能、认知、心理等,这些不同术语都被无区别地使用)的测试。如果机器通过了测试(赢得游戏),就足以认定它能够“思维”,而且如果计算机能经由编程通过测试,那么思维的本质就是计算。如果思维只是心理算法的运行(还有其他的可能吗?),那么,丘奇-图灵论题告诉我们,这些心理算法都可以由图灵机来完成。图灵定理告诉我们,所有的这些机器都可以由普适图灵机来模拟。因此,我们就可以推论,思维是计算的一个种类,类似于在普适图灵机上的计算,例如冯·诺伊曼机或者产生式系统。由此看来,人工智能、认知心理学还有神经科学的工作,就是要找出心智的真正算法(下一章会给出更多论证)。

185

普特南与图灵机器功能主义

1960 年,普特南(Hillary Putnam)出版了《心灵与机器》一书,探讨了与心—身难题相关的心灵和图灵机的**类比**问题。但我们还不清楚他是否有意提出心智的计算理论。例如,他在书中评论说:“我并没有说这个观点(即是否能够识别心理事件和物理事件)是针对图灵机而提出的。”但如果心理事件是计算过程的话,有人可能会问:为什么不呢?后来普特南评述道:“我可能因为强调这种类比(心灵与图灵机),而被人指责鼓吹一种机械主义的世界观。但如果因此就认为我赞同机器会思维,此其一;或者人是机器,此其二,那么这样的理解都是错误的。”这听起来不像是心智的计算理论——心理状态等同于计算状态。在同一篇文章中,普特南明确指出图灵机可以有多种实现方式,即“普适图灵机是一种抽象机器,它在物理上几乎有无数种可实现方式”。但重要的是,他对于心理状态并未持相同观点。这一点在他 1967 年的论文《心理状态的本质》(原文的标题是《心理指谓》)里表述得很清楚,这篇文章很可能就是心智计算理论的滥觞。论文(Putnam, 1967)中还讨论了几个重要步骤:首先,把心理状态等同于功能状态:“简而言之,我认

为**疼痛**并不是一种大脑的状态，即不是大脑（或者甚至也不是整个神经系统）的物理一化学状态，而是整个有机体的**功能**状态。”第二，他进一步阐明了心理状态可有多种可实现方式的观点：“哺乳动物的大脑、爬行动物的大脑、软体动物的大脑……乃至任何地球以外可以找到的所有动物都能感觉疼痛……”由此可知，(1)图灵机可有多种实现方式；(2)心理状态可有多种实现方式；(3)心理状态是指特定的功能状态。至此，普特南认为，对图灵机和心理状态两者共同具有多种可实现方式的**最佳解释**是：图灵机的不同状态指定了相关的功能状态，同时能够对应不同的心理状态。当然也许有人会（严厉地）批评说，普特南并没有谈及任何有关**认知**的东西，只是谈到“疼痛”而已。但是，有理由认为，普特南只是把疼痛当作心理状态的一个典型例子，因而他的论据可以被推论到**一般的心理状态**上。例如，他在反对“心物同一论”时，说道：“心物同一论者认为不仅**疼痛**是一种脑状态，而且**任何**心理状态都是脑状态。”显然，如果图灵机器功能主义要反驳心物同一论，它也必须是关于所有心理状态的理论（强调它的普适性）：“如果一个程序……成功了，那么紧跟着就会……带来对‘心理状态’这个概念的准确的定义。”我们知道，这篇论文的标题不是《**疼痛**状态的本质》（也不是《**疼痛**的指谓》），而是《**心理**状态的本质》——普适性是重要的。因此，与图灵相比，更有可能是普特南跨出了从“计算的智能”到“智能的计算理论”关键的一步。 186

布洛克和福多

布洛克和福多(Block and Fodor, 1972)把普特南的图灵机（或概率自动机）思想扩展到了一般的计算系统，最后完成了对心智计算理论的建构（福多对CTM进行了扩展和提炼，讨论了下列理论和问题：思维语言(1975)、模块性(1983)、常识心理学与命题态度分析(1987)、狭义内容与意向归纳(1994)。

纽厄尔和西蒙

到目前为止，我们只是在哲学范围内回顾了CTM的历史，但从更广义的认知科学意义上来看，还有其他领域对其产生了影响，特别是人工智能(AI)。在这方面最有影响的成果之一是纽厄尔和西蒙(Newell and Simon, 1976)的论文，他们在文中提出了：

> *物理符号系统假设*：物理符号系统具有产生智能行为的充分必要条件。

可见，这个观点认为系统具有产生“智能行为”的能力，而没说它直接具有认知或者心灵，因而人们会说机械系统的行为是“智能的”，而不会说它们

具有心智，就像没有人会说“智能炸弹”具有心智一样。但毫无疑问的是，这个假设试图表明，智能的先决条件是认知或者心智，似乎已成为本领域的一
187 条理论预设。但什么是“物理符号系统”？它是指储存在机器里的一组物理符号（例如，在硅片或者神经元上实现的符号）和操作：“物理符号系统包含一组称为符号的实体，这些实体是某种物理模式，是称之为表达式（或者符号结构）的另一实体的组成部分……除了这些结构之外，系统还包含一组操作程序，能够作用于表达式使其产生另外的表达式……**物理符号系统就是在机器内能够产生出随时间而演化发展的符号结构集合**”（Newell and Simon, 1976）。既然这里所谈论的“机器”等同于普适图灵机（UTM），这个假设认为机器拥有了智能行为的能力等同于图灵机的实现。

心智的计算理论

在这些历史述评之后，我们转向目前人们阐述的一般心智计算理论（CTM）的普遍形式（采用福多对 CTM 的表述，Fodor，1987）。CTM 对认知的普遍理解，尤其是对命题态度的心理状态的理解，现在都可以看作是 RTM 的特殊个案。CTM 试图对所涉及到的关系、操作和表征的本质作出更为详细的规定。

1. 在 RTM 中，**关系**是计算的。
2. 在 RTM 中，**操作**是计算的。
3. 在 RTM 中，**表征**是计算的。

我们首先讨论前面两点，第三点放在下一节讨论。

形式约束

关于计算的关系和操作，福多（Fodor, 1981b）认为心智的计算理论必须遵守他所谓的“形式约束”：“我认为心智计算的过程是……符号的，因为它们是通过表征来定义的，同时它们又是形式化的，因为它们（概略地）借助于表征的句法操作表征……形式操作不需要通过诸如真值、指称和意指等表征语义特征加以说明……形式操作是指对特定范围内表征对象的形态进行的操作……按照心智的计算理论，两种不同观念，只有在当它们能够被识别出不同的形式表征关系时，才能够区分”（1981b：227）。当我们想到数字计算机的编程意味着什么时，形式约束就显得充分合理了。编写的程序（TM
188 指令表、vNM 程序、PS 产生式）直接对数据结构的格式或形态进行操作，而不是数据结构所指代的对象。

术语解注 福多在论述形式约束的时候，经常不加区别地使用“形式”和“句法”。已经有人，例如戴维特（Devitt，1989，1991）就强烈地抱怨说，“形式”与事物的内在形态（格式）有关，而“句法”则与表征之间的外部关系有关。虽然我们也认同这种区别是重要的和恰当的，但似乎未引起学界的重视（弗雷格和罗素曾谈到“逻辑形式”，而卡尔纳普则用“逻辑句法”）。我们把“句法”当作是“形式”的一种格式变体。

态度

形式约束告诉我们，什么时候会有两种不同表征，而在什么时候会产生两种不同态度？例如，如何区别相信、意愿和计划？考虑一下空想先生的例子（见图 8.2）。我们知道在诸如“相信”和“意愿”的标签下具有不同的表征集合。决定一个表征属于相信而不是意愿的，是这个表征与其他表征之间的交互方式——即在理论和行为动机推理中它被利用和操作的方式。如果你**想要**喝点什么并且相信冰箱里有可乐，在其他条件都满足的情况下，就会**计划**走到冰箱前取可乐喝。这通常要涉及“相信盒”中构成相信的表征之间的一系列互动，以及“意愿盒”中构成意愿的表征之间的一系列互动，其他命题态度也是如此（又见图 8.2）。这里的“盒”是为了将复杂的计算关系形象化而作的简化处理。这些计算的关系和操作尽管一般是通过编程的方式输入机器的，但有时却是硬线连接的（我们将在本书的结尾部分回到这两者的区别上）。目前我们仍使用这个中性的术语。

189

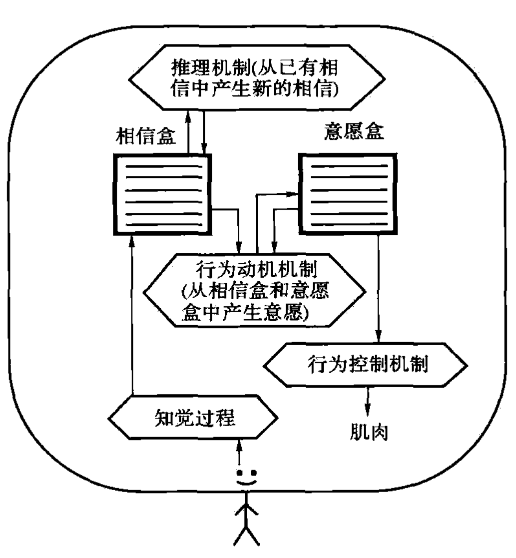


图 8.2 空想先生（改自：Stich，1983：75；经 MIT Press 许可复制）

表征也有语义,或者说,表征具有"心理内容",因为它们表达了命题,并因此而指向某事物。我们可以把一般 CTM 表述为:

(CTM)

1. 认知状态是具有内容的**计算心理表征的计算关系**。

2. 认知过程(认知状态的改变)是具有内容的**计算心理表征的计算操作**。

到目前为止,我们已经有了一种一般心智计算理论。但如何才能得出明确的心智的**数字**计算理论呢? 这是目前某些争论的焦点(后文涉及)。现在我们只是讨论两个合理条件:

结构 心智的数字计算理论附加的一个约束是,**计算结构(记忆和控制)必须是数字的**。

表征 心智的数字计算理论所附加的另外一个约束是,**表征必须是数字的**,也就是称之为数字计算理论的原因。

根据这两个条件,可以用 CTM 的数字规定表述基本 DCTM 如下:

(B-DCTM)

190

1. 认知状态是具有内容的**计算心理表征的计算关系**。

2. 认知过程(认知状态的改变)是内容的**计算心理表征的计算操作**。

3. 计算的结构和表征(上述第 1 点和第 2 点)都是**数字的**。

最后一点,还必须知道,每一种命题态度,例如"相信",都是一种概念束(包含多种不同的相信),因此在相信集合中就会包含多种计算关系。

8.3 心智的数字计算理论与思维语言

还需要强调,CTM 把认知状态和过程分解为**两个部分**:一部分是计算关系或者操作,另一部分是心理表征。数字机器的一个普遍特征是,它们均以某种编码方式进行计算。如果心智计算是一种数字计算,那么,心智对心理表征的操作(生成、转译和删除)和心理表征自身实际上就形成了大脑的一种机器编码。在某种程度上,这种系统编码的结构意味着心理表征形成了一种类似于语言的系统,这就是所谓的思维语言(LOT)**假设**(见 Fodor,1975)。把这个观点与 CTM 整合到一起,就得到了第一种正式的心智数字

计算理论：

(DCTM)

1. 认知状态是具有内容的**计算心理表征**（在思维语言中）**的计算关系**。

2. 认知过程（认知状态的改变）是具有内容的**计算心理表征**（在思维语言中）**的计算操作**。

3. 计算的结构和表征（上述第 1 点和第 2 点）都是**数字的**。

某种程度上，是存在思维语言证据的，思维语言能够对 DCTM 中某一方面提供支持。那么什么是 LOT？又有什么理由接受它呢？

思维语言

思维语言（LOT）假设背后的基本观点是，认知的表征系统**与语言类似**。191
显然，自然口语有一些特征与认知无关，例如聊天、询问、倾听等，因此思维语言在一些方面与（口语）自然语言并不相似。但在哪些方面相似呢？典型的有：

(LOT)

1. 思维语言系统有一个关于表征的基本"词汇表"，包括概念、感知和图像等。

2. 词汇表中的条目，可根据普遍原则进行组合——这就是说，系统具有"句法"。

3. 条目以及它们之间的结构性组合是**指涉**某事物的——具有组合"语义"。

福多最初的观点认为，这样的系统类似于计算机的"机器语言"，机器语言是机器设计的一部分，所以机器并不需要学习。以此类推，LOT 是我们先天基因的一部分，必须被看成是在我们学习第一语言之前用于思考的系统。当然，这个系统的成熟可能需要一定时间，因此在出生时它还并未完整（类似于牙齿或者头发一样）。LOT 假设的这种"先天"观点是极具争议性的，在后面的讨论中我们将不采用这个观点。

这样一个系统的存在以及它的本质如何对 DCTM 产生影响呢？思维语言对于 DCTM 的影响大概是这样的：在计划和推理的时候，人有一个用于表征他们周围世界的心理系统，计算机也如此。此外，心理表征系统还具有生成和系统特征，计算机也如此。人和计算机进一步的类比支持这样的观点：

人类心理生活的各个方面都可以看作是计算。下面更详细地探讨这两个方面的理据。

LOT 的理据:呈现模式(表征方式)

这是一个(非常)简短的故事:爱丽丝和贝蒂正在翻阅她们 1959 年的中学年鉴,这所中学位于明尼苏达州希宾市。翻到最后一页,爱丽丝说:“齐默曼,记得他么,罗伯特・齐默曼?皮包骨头、头发长长的家伙,高中毕业后就离开了城市。听说他上过明尼苏达大学,但是没有毕业,离开了城市,不知道他后来怎么样了?”这时,她们瞥了一眼时钟,六点半了,准备出发的时间到了,她们的回忆随之中断了。三个月前,她们购买了鲍勃・迪安的音乐会的门票,八点钟开始。故事就这样结束了。这种情景会发生么?好像会。但是,如果鲍勃・迪安**就是**(那位)罗伯特・齐默曼的话,这种情景还会发生么?如果会,她们又是如何在想知道他后来怎样了的同时,又知道她们将去听谁的音乐会呢?有人会说,啊,是的,但她们并没有**认识到**,她们一直在阅读的罗伯特・齐默曼和正要去听的音乐会的主角鲍勃・迪安,与她们高中时候反社会的那个同学正是同一个人。如何解释这种可能性?根据 LOT,这个故事让我们有理由相信,爱丽丝、贝蒂以及大多数人,都是应用了指向外界事物的,并以某种特定方式表达的表征来思考事物的——表征是只针对心智自身“描绘”或者“说明”外部事物的方式——这样所引发的一个结果就是,人们往往无法辨认以两种不同方式所表征的同一事物,但在“口语系统”中同一对象却可以有多种表达。这就是说,我们仍然没有理由认为表征在句法结构和语义组合上与“口语”系统类似。下面简要讨论这种观点的依据。

192

LOT 的理据:产生性

LOT 的产生性来源于这样一个事实,即人的思维是产生性的,能够依据有限已知推出独特而新奇的无限序列,例如,已知 1 是一个数字,2 是一个数字,就可以无限类推。这种能力似乎不受任何思维本身的特征限制,而主要取决于一些外在因素的影响,例如动机和注意、经验等。然而,人的生理特征是有局限性的,具有有限的记忆和操作,那么具有这种生理有限性的人,是如何具有产生潜在无限序列观念的能力呢?LOT 认为,心智的有限表征系统带有一种组合结构,LOT 通过这种组合结构将旧的碎片拼凑成新的组合。

LOT 的理据:系统性

这个 LOT 假设的理据是它具有形式化特征。人类(以及某些动物)的认知具有一些只有 LOT 假设才能解释的系统性特性,而其他理论对此无法说明。因此依据选择最优理论的原则,自然就得出这个结论:LOT 假设是真的。

思维

最初,福多用 LOT 解释思维的系统性作为 LOT 成立的依据,他说道: 193 "LOT 认为,人拥有某种观念与表征的排列结构相关,那么,'约翰爱玛丽'和'玛丽爱约翰'这两种观念都涉及同样的表征和同样的表征结构。因此,自然地,任何人如果有与其他人具有相同的观念,就必然拥有与其他人相同的表征和表征结构。这样 LOT 就解释了思维的系统性"(1987:151)。

推论

第二个支持 LOT 的相关提议是由福多和派利夏恩(Fodor and Pylyshyn,1988)提出的,这个提议是以**逻辑系统性**为基础的:"所有的心智理论(包括 LOT)都要遵循下面这个基本预设:理论中所包含的模糊逻辑类型的心理推论,应当在相当程度上符合认知的特点。例如,似乎不可能在现实生活中找到这样的心理状态:从 P&Q&R 能够推出 P,却不能从 P&Q 推出 P(1988:46-7)。

正如这些作者指出的那样,心智几乎不可能出现下面的情况:心智中包含着某一类别的观念或推论,但这些观念或推论却与其他的观念或推论毫不相关——"具点心智(punctate minds)"是不存在的,比如,含有 73 种互不相关观念的心智是不存在的。以上两个方面的理据说明,思维和心理推理实际上具有系统的特征。LOT 假设认为,思维和心理推理在表征的类语言系统中执行。如果 LOT 假设是真的话,那么,思维和心理推理的系统性就能得到解释。到目前为止,还没有其他更好的理论能解释思维和心理推理的系统性,所以 LOT 假设可能是真的。

这些支持 LOT 的理由(呈现模式、产生性以及系统性)与支持系统性的经验证据一样,都是有效的,但是,某些联结主义者却认为仍需严格的证明。

8.4 DCTM与心一身问题

DCTM可以进一步阐述为:(1)它为一个传统的哲学难题即"心一身问题"提供了迄今为止最好的解决方案;(2)它为认知的实证研究提供了一个富有成效的框架。这是其他心智理论所不能比拟的,值得我们认真探究。

194

心一身问题

心一身问题可以表述为:心理现象(状态、事件和过程)与生理现象(状态、事件和过程)之间是什么关系? DCTM的优点之一是,它提供了建构成功的认知理论的一个框架,部分原因是:(1)它提出了解释心一身问题的方案,且未被与之竞争的理论所驳倒;(2)它使对行为的**心理一逻辑解释**变成一种可接受的**因果解释**。我们将逐一讨论,先考察与之竞争的相关理论,再指出它们各自的缺点。

竞争理论

二元交互论

世界可分为心理和物质**两种不同的实体**,这种观点因笛卡尔而为人所熟知(见第3章)。心理实体有意识和时间维度但没有空间维度(因而不可分割,不受物理法则的约束),而物质实体则既有时间维度也有空间维度(因而是可分割的),所以受物理法则的约束。这些实体之间能够产生**因果关系**——心理事件能导致生理事件,同时生理事件也能产生心理事件。

支持理由

身心之间的**交互作用**与我们的直觉相符:生理事件(嗜酒过度)会导致心理反应(醉),心理原因(决定要挥手再见)能引起身体反应(挥手再见)。

反对理由

1. **身心二元**的观点与实验心理学相矛盾,后者包括很多物理科学的方法,而物理科学方法似乎都不能应用于心理实体(见上)。

2. **身心的因果关系**难以解释——在不违反能量守恒等定律的情况下,没有空间和物质的实体如何作用于具有空间和物质的实体?

副现象论

这种观点认为，虽然生理现象可以引起心理现象，但是心理现象却不能影响生理状态。据此观点，计算机的闪光信号灯（或者发热）是它运行的副产品，对其本身运行并没有帮助。同理，心理现象只是大脑工作的副产品，不会对大脑的运行有什么作用。 195

支持理由

1. 避免了上述二元论的第二个缺点——例如，不会违反能量和动量守恒定律。

反对理由

1. 不能避免上述二元论的第一个缺点。

2. 它使人类的思维、判断等心理活动与人类的历史进程无关，这一点难以让人接受。

激进行为主义

这种观点认为，心理现象就是特定类型的行为——对特定刺激作出的特定反应。

支持理由

1. 这种观点通过**消解**心理现象"解决"了心一身问题——没有了心理现象，也就没有了心和身之间的关系问题。

反对理由

1. 如果没有足够的证据，难以让人相信心理现象是不存在的，以及心一身之间没有任何"交互"的观点。

2. 心理学并没有完全回避心理现象存在的问题。当代认知心理学已经有力地证明了心理结构可以产生行为。

逻辑（分析）行为主义

这是一种语义的观点，认为每一种解释心理现象的陈述都**等价于**某一种行为假设（如果有某事发生，那么就会导致某种结果），这如同对玻璃杯易碎的解释，可以用如果玻璃杯掉在地上就会破碎等诸如此类的陈述进行分析。

支持理由

1. 它把心理的因果关系描述为行为假设（产生某种结果的原因：如果某事发生，那么就会产生这种结果，并且某事确实发生了）。 196

反对理由

1. 没有理由相信，行为假设能够解释所有的心理现象——还不存在这样一种行为假设，它能够充分解释某一**心理现象**。

2. 它不能解释（心理）事件的因果关系：是感到口渴并注意到桌子上有一杯水，才会伸手去拿那个杯子，而不是反过来。这是一种基本的因果关系。

物理主义

（由于物理主义假设心—物的同一性，也称为**同一性理论**。）同一性理论有弱和强两种文本。

殊型物理主义

这种观点认为心理现象的**殊型（个例）**和物理现象的**殊型（个例）**等同，但是心理的和物理的类型未必相同。假设对人来说，每种实际的疼痛，都是对一种叫做C-纤维的刺激，而C-纤维最终是由碳氢化合物构成的。再假设火星人是由硅构成的，当他们感到疼痛的时候，则是S-纤维受到刺激。这种对S-纤维的刺激与对C-纤维的刺激是不同类型的物理现象，因而也是不同类型的心理现象。

类型物理主义

这种观点认为**心理现象的类型**和某些**物理现象的类型**等同。据此，当两个系统表现出同一类型的物理现象时（例如，对C-纤维的刺激），它们就能表现出相同类型的心理现象（例如，疼痛）。因此，既然硅和碳氢化合物不同，且如果疼痛是一种C-纤维刺激，而C-纤维是碳氢化合物的话，那么由硅构成的火星人和由碳氢化合物构成的人类就不可能感到同样的疼痛。类型物理主义者同时也是殊型物理主义者（既然所有的心理现象的类型等同于物理现象的类型，那么心理现象的个例当然也等同于物理现象的个例）。类型物理主义是一种高度还原的理论——它把心理现象**还原**为物理现象，但与“取消论”不同，它没有否认心理现象的存在（例如，如果鲍勃·迪安就是罗伯特·齐默曼，并没有**取消**鲍勃·迪安——鲍勃·迪安仍然存在）。

197

支持殊型物理主义的理由

1. 它提供了一种对心理事件的解释，因为每个特殊的心理事件**是**（等同于）某种物理事件，所以“心理的”因果关系就是物理因果关系的一个种类。

支持类型物理主义的理由

1. 除了具有上述殊型物理主义的优点之外，还为心理概念提供了一个指称对象。例如，疼痛是对C-纤维的激活——回答了这样一个问题：系统必须在什么状态下才能感到诸如疼痛的感觉？

反对殊型物理主义的理由

1. 没有为诸如疼痛这样的心理概念提供一个指称对象，即没有回答这个问题：系统必须在什么状态下才能感到诸如疼痛的感觉？

反对类型物理主义的理由

1. 它把心理现象的类型等同于物理现象的类型，但心理现象似乎更多依靠“软件”而不是“硬件”。假定由硅构成的火星人有适当的软件支持的话，他们也有可能感到疼痛或相信地球是圆的。

2. 心理（或者至少是认知的，我们后面将回到这个主题）现象似乎一般是指**加工信息的系统**，而不是讨论系统由什么东西构成。

评价

殊型物理主义心理现象和物理现象的同一太弱，不能让人满意，而类型物理主义则过强。我们是否可以在拥有同一性理论优点的同时，又能够避免它的缺陷呢？

随附性

一种被人熟知的且具有同一性优点的观点是“随附性”理论，把它与功能主义（见后文）相结合，是目前解释心一身关系最流行的方案之一——因此我们会用较长篇幅探讨这个观点。那么，什么是随附性呢？我们先来看一些心一身问题之外的例子。所有的状态都会伴随着发生其他状态，摩尔(G. E. Moore)从艺术角度引入了这个概念。以油画《蒙娜丽莎》为例，我们用“分子对分子”的方式对它加以复制——形成两个孪生的《蒙娜丽莎》。第一幅油画很美，第二幅油画有可能不美吗？由于它们在颜料分子的水平上是相同的，所以使第一幅油画产生美的原因，也同样符合第二幅油画。因而，对这一问题的回答是“否”，因为我们说油画的美**随附**于它的物理特征（颜料分子）。再想一下秃顶的例子——如果一个人秃顶了，他和他的孪生兄弟“分子对分子”（同样地，“头发对头发”）相同，他的孪生兄弟有可能不会秃顶吗？答案同样是“否”。分子水平相同的孪生兄弟具有相同的头发分布，他们同样会秃顶，因为秃顶**随附**于物理特性（头发的分布）。所以，在更一般的意义上，我们说：

198

一般随附性（generic supervenience，GS）：如果不存在这样的两种情形，就B-事实（油画、头发）而言，它们是相同的，但就A-事实（美、秃顶）而言是不同的，那么A-事实**随附**B-事实。或者这么说：如果存在这样两种情形，就B-事实（油画、头发）而言它们相同，那么它们就A-事实而言也相同的必要条件是，A-事实（美、秃顶）随附于B-事实。

现在我们再回到心身问题的探讨，这一问题用随附性来解释更富有争议性。

一般心一身随附性（generic mind-body supervenience，GMBS）：如果不存在这样的两种情形，就B-事实（生理的）而言它们是相同的，但就A-事实（心理的）而言是不同的，那么心理事实（状态和加工过程）**随附**于生理事实（状态和加工过程）。

在随附性的特征以及上述例子中蕴含着如下三种观点。第一，"高层次"的特性（美、秃顶、心理体验）与"低层次"的物理特性相**协变**（covary）——两层次属性之间具有一些"共同点"和"不同点"。第二，高层次的特性以某种方式**依赖**于低层次的物理特性，物理事实以这种方式确定或影响高层次事实——确定了物理事实，同时也就确定了其他事实。第三，高层次特性**不能还原**为低层次的物理特性——这种特征与类型物理主义不同，它认为高层次特性是与低层次特性截然不同的另一种特性。这种论点有时被称为**属性二元论**。它在具有类型物理主义优点的同时，避免了还原主义的疑问。下面我们深入辨识一些概念，会得到一些关于随附性的新认识。一般而言，GS和GMBS存在两种分歧：(1)产生作用的可能必要条件是什么；(2)关于什么样的情形的争论。

必要条件的可能性：其中可能蕴含这样的观点，B-事实通过逻辑必要条件（"逻辑随附性"），或者自然法则（自然随附性）决定A-事实。逻辑和自然法则是不同的，例如，光速（186,000英里/秒）是最快的，这是自然法则，但在逻辑上我们还是可以说某种物质的速度比光速还快1英里/秒——也许宇宙里有不同的物理法则。

199

情境：我们可能会说，测试A-事实和B-事实的相同和不同点是"局部的"，只限于单个个体——如《蒙娜丽莎》和男人秃顶的例子。对于单个个体而言，如果复制了B-事实就会复制A-事实的话，那么A-事实"局部随附于"B-事实。进一步，我们推论对于整体情境而言随附的情形，即个体所嵌入的整体情境——它只是整体中的一个成员。对于整体而言，如果复制了B-事实就复制A-事实的话，那么A-事实"整体随附于"B-事实。在A-事实涉及与系统所在世界的关系时（与前面的美和秃顶的例子不同），局部和整体随附

性的不同就变得很重要了。例如，两个生理结构相同的生物物种，可能因为他们生存环境的不同，而具有不同的生存或者“适应”价值。因此，根据达尔文理论，生物的适应特性不是局部随附于他们的生理结构，而是整体的随附。因此随附性理论需要包含生物物种所生存的环境，才能保持整个物种世界的同一性。

对于心身之间是哪一种随附性关系，存在着不同观点：逻辑的还是自然的，局部还是整体？但无论选择哪一个，都会有所漏失；对于下面这个问题，随附性理论不能像考察过的那些理论一样，能够给予完整回答：“心理现象是什么?”随附性提供了一种心和身之间的复杂关系（在这个意义上，它回答了：身心的关系**是**什么?），但没有给出心理究竟是什么的洞见——毕竟，美附着于颜料上和秃顶附着于头发上并不能说明什么是美和秃顶。下一节中，也是最后一种要介绍的理论，将试图涵盖随附性的优点，并填补这个缺陷。

功能主义

一般而言，功能主义理论通过功能，也就是看它们能做什么，来识别事物。例如，捕鼠器是能捕捉老鼠的东西，汽化器是把汽油和空气按易燃的比例混合后送进气缸的东西。捕鼠器和汽化器都不是以它们的**构成**来定义的。事实上，这类事物只要具备充分的结构，能够实现相应功能，就可以是由任何材料构成的——木材、塑料、铜、钢铁等等。因此，如果说存在什么捕鼠器和汽化器的“类型物理”，就肯定是错误的。实现捕鼠器和汽化器的方式是“可多种实现的”——它们都可以用多种材料实现。同理，功能主义的心智理论宣称，心理现象必须通过它们的功能加以辨别。然而，心理现象的功能是什么？它们能做什么呢？ 200

在这里，功能主义的心智理论已经多少偏离了通常意义上的有什么用途的含义（通常意义上的“功能”是指系统能做什么，例如捕捉老鼠）。例如，功能主义并没有说，相信是具有某种功能的东西，就像捕鼠器的例子一样。而是说，功能主义的心理“功能”是指，在**认知系统中产生某种作用的一种表征**。再想一下空想先生的例子，他的相信和意愿被描绘为包含相信和意愿内容表征（他所相信的**东西**和他所渴望的**东西**）的盒子。但这种盒子画像（以及盒子之间的交流），只是对表征内容在系统中所起到的复杂作用——表征与表征之间以及表征与系统本身之间的复杂关系——而作的一种简便式虚构速写。功能主义认为这些复杂关系（作用）包括：**输入**关系、**输出**关系，以及它们与其他**内在**状态和过程之间的关系。

功能主义和心一身问题

功能主义赞同殊型物理主义，把殊型或者个例的状态、事件和过程与**类型**或者**种类**的状态、事件和过程区分开来。**纯粹功能主义**在回答心一身问题时认为，每种心理状态、事件和过程的**类型**与系统中某些功能状态、事件和过程的**类型**等同，纯粹功能主义对系统由什么材料构成的这一问题是开放的。**物理功能主义**认为每种心理状态、事件和过程的**殊型**与其物理状态、事件和过程的**殊型**是**等同**的。因此，根据物理功能主义（与殊型物理主义一样），不存在随意发生的心理现象——所有的心理现象最终都等同于物理现象。

术语解注 现今所有的功能主义者实际上都是物理功能主义者，因此除非明确提及它们的区别，我们将用“功能主义”指代“物理功能主义”。

支持理由

功能主义不但同时具有二元论、行为主义和物理主义的优点，而且又避开了它们的缺点。第一，与**行为主义**不同，而与**二元论**相似，功能主义认为心理因果关系是真实因果关系的一种类型，因为它把“心理”现象的殊型等同于物理现象的殊型——殊型的心理事件能够产生殊型的物理（或者其他心理）事件——而**二元论**则把它看成是一种神秘关系[1]。第二，功能主义从关系的角度来定义心理现象，这一点与**行为主义**相同；功能主义允许使用**内部**状态之间的关系，以及它们与输入和输出状态之间的关系来定义心理现象，这一点又与行为主义不同。第三，**类型物理主义**把心理类型等同于物理类型，因此似乎否定了认知的**多种可实现性**，而功能主义却与之不同，允许这种可能性。也就是说，如果心理状态的类型是物理状态的类型，那么，任何具有心理现象的系统就必然共有一些物理特性。但很多人相信，没有理由认为认知系统不能由其他各类神经组织或者硅构成。最后一点，又与**殊型物理主义**不同，功能主义描述了心理现象的类型——它能够回答这个问题：系统必须具备什么样的特性，才能够表现出相同的心理现象？

机器功能主义与 DCTM

既然在一个复杂系统里，可能存在多种作用/关系能够定义心理状态，那么，就需要对如何从这些可能中进行选择给予约束——是什么样的约束呢？

(一般)机器功能主义

机器功能主义并不绑定任何一种特殊的计算构造,虽然作为历史事实,机器功能主义从图灵机(虽然冯·诺依曼机可能更合适)获益最多。任何"机器",只要具备输入、内部和输出的关系就符合功能主义的要求,因为机器功能主义能够允许只有三种类型的作用/关系作为其功能状态。因此,心理现象也可以同样描述为**输入**作用(roles)/关系、**内部**作用/关系和**输出**作用/关系。根据机器功能主义,**心理状态(过程等)的每一种类型都等同于机器功能状态(过程等)的某一种类型**,而功能状态(过程等)的每一种类型都可以用系统的输入、内部和输出的作用/关系来定义。

DCTM

如果把机器功能主义中的**功能关系**,转译为 DCTM 中提到的**计算关系**(当然,仍保留 LOT 中的表征),就从机器功能主义得到了心智的数字计算理论(DCTM)。可见,DCTM 是机器功能主义的一个特例,而机器功能主义又是功能主义的一个特例。因为功能主义已经对心一身问题作出了最好的回答,所以 DCTM 自然地从功能主义那里继承了这个优点,这点也是对DCTM 给出的最后一个理据。 202

功能主义、随附性和物理主义

如果心理状态(包括过程)的类型是功能状态的类型,而心理状态的殊型是物理状态的殊型,那么功能和物理的关系是怎样的呢?一种标准的回答是,功能状态"随附于"物理状态。也就是说,如果两个系统的所有物理特性(状态)相同,它们的功能特性(状态)也必然相同。如果用系统能够做什么定义功能状态,例如系统能导致什么事情的发生,那么功能状态"随附于"物理状态的观点就更加清晰了。物理上相同的系统当然具有相同的因果作用关系,一个系统**能做**的事另外一个也能做,因此物理上相同的系统应该具有相同的功能状态。以捕鼠器为例,两个"分子对分子"的孪生捕鼠器,毫无疑问,如果一个能捕捉老鼠的话,那么另外一个当然也能。

再回到心理的问题上。如果心理现象只是功能现象,而功能现象随附于物理现象,那么心理现象也就随附于物理现象。这意味着,物理上的孪生会具有相同的心理状态(孪生兄弟会进行同样的思考)。这个观点有些过于激进,我们将在后面继续讨论这个问题(见第 9 章)。

8.5 DCTM与表征内容

DCTM对心理现象提供了一种因果解释,因而提供了我们希望从心智理论中获得的那部分内容。那么它又是如何处理内容难题的呢?在前面的章节,已经提到两种不同表征类型的系统在“语义”上各自的优点和缺陷,这两种系统分别是青蛙和数字计算机。

昆虫探测器与数据结构

青蛙

青蛙视觉系统的优势,是能够表征(探测和追踪)特定的目标及周围环境,虽然探测和追踪的范围有限。我们在第4章为这个系统描述了一个内容(关于语义)理论——“简单探测器语义”(SDS)。按照SDS,青蛙视觉系统的缺陷,是很难对它会错误表征目标及周围环境作出解释,或者很难解释它为什么会只表征特定的目标及环境。此外,SDS不能很好地说明表征之间的逻辑关系,因此很难认为它反映了表征的全貌。

203

数字计算机

另一方面,数字计算机却能很好地解决表征之间逻辑关系的问题,但它的难题是如何做到合理地探测或者追踪特定目标及周围环境?回顾(第7章)关于数字表征如何进行表征的唯一方案——连接表征与所表征对象——是解释语义(IS)。据此,表征关系解释为表征与所表征对象的同构性。但我们知道,仅用同构性来解释并不令人满意。

更贴切地说,“机器思维”的内容似乎是由程序员(可以广泛地理解为所有解释机器输入和输出的群体)所决定的。例如,如果程序员给机器加载一个程序,告诉它移动棋子主教到棋盘的某个位置,我们会认为是机器在解决一个象棋问题。如果程序员加载一个程序,告诉机器把坦克营移动到某个地理位置,我们会认为是机器在计划一场沙漠战役。可以这样理解,把棋子主教移动到某一位置与把坦克营移动到某一位置是同构的,形式上是相同的程序——只是对程序的解释(**对人而不是对机器而言**)不同而已。那么是什么使机器能够“思维”象棋和战役呢?这个问题不能用计算来回答,而需

要引入与环境的某种联系。

数字计算的正式观点:功能作用

前面已讨论过,功能主义使用三种关系分析认知现象:输入、内部和输出。例如,有这样一种相信:**那个男人很高**,符合表征内容的关系是什么呢?有人(见 Fodor,1981a)提供了下面的解释:

输入 看到这个男人的身高
内部 据之推论出相信:**某人很高**
输出 说出"那个男人很高"

根据这种观点,相信的内容是由所有涉及的相互关系确定或决定的,即它的**功能作用**。这样它的优势之一,就是同时具备了青蛙"昆虫探测器"(与环境的连接关系)和数据结构(逻辑关系)的优点。为了说明这点,再回到起初的例子:**那个男人很高**: 204

1. **输入**是指某个表征的出现与某个特定男人的联系,类似于青蛙纤维-2 的激活对应着附近一只昆虫的出现。而**输出**是指表征与最终语句之间的联系,类似于青蛙纤维-2 的激活对应着捕捉昆虫事件的发生。

2. 得出**某人很高**推理的**内部**关系与数字表征间的蕴涵关系(见第 7 章)相同。

因此,根据这种观点,心理表征的内容取决于两个因素(或者方面)——一个是内部因素(内部功能或者"概念"作用),即一种表征与其他表征的相互关系;另一个是外部因素(外部功能作用),即表征与所表征的对象以及周围环境的相互关系,心理表征的内容因此成为心理指称以及各种心理状态真假的基础。布洛克(Block,1986)把这两个因素分别称之为"短距"和"长距"概念或功能作用。不过,我们认为"概念作用(conceptual role)"这个术语仅适合短距内部因素,那么它们的关系可精确地表述为:**概念作用**=内部功能作用="短距"概念性作用;**外部功能作用**="长距"概念角色。

DCTM 与功能作用

"长距"功能作用如何进行**计算**——外部输入和输出的功能关系?用功能主义分析心智,这种外部功能关系经常使用诸如感知(输入)和行为(输出)这些现象进行说明。然而,用功能主义分析图灵机,它的输入和输出关

系是在磁带上读写，但磁带本身又是机器的**存储**，并不属于周围环境，不在机器"外部"。因此，(机器)功能主义的这个范例就把表征内容限定在内部概念作用上。那么，DCTM能够使用功能主义的另一部分——以外部、长距的功能作用方式，如感知和行为——进行说明吗？还是说DCTM只能用内部概念作用表述？

205 如果要遵循**形式约束**，那么信息在计算机中的表征，就只能依据形式约束进行操作，这就强烈地意味着表征内容仅由内部、概念作用决定。如果这一说法成立，那么DCTM在某种意义上要比功能主义狭窄，因为功能主义允许外部功能作用同样能够决定思维内容。另一方面，DCTM承认心理的状态和过程是具有内容的心理表征的计算关系和操作，但关于表征内容的外部因素如何确定，并不是DCTM的一部分，DCTM仅仅关注内部概念作用如何确定。同样，DCTM宣称认知状态和过程是表征(数据结构)的计算关系和对表征的操作。这些结构**具有**内容——它们表征某事物。但是，(1)它们所表征的事物在计算中并不起直接作用，只提供表征的"形式"、"句法"和"结构"；(2)它们所表征的事物并非(总是)可以用计算术语解释清楚。因而在某种意义上，DCTM对表征内容的前提或者预设(倾向于)采取一种非计算的解释。暂且认为就是如此，我们再返回到计算问题的讨论上：什么是概念作用，它由什么决定呢？

概念作用

对概念作用(conceptual role, CR)的一般理解，可称之为"一般"CR理论：

> **(G-CR)**
> 表征内容是由该表征在其所处的概念系统中的作用所决定的。

对表征本身，以及它们在所属概念系统中扮演的作用和其所属的概念系统有不同理解，会得到不同的CR理论。例如，把CR和条件或然性联系起来(参见Field，1977：390)，是早期的一种重要观点：

> **(CP-CR)**
> 如果一个表征的条件或然性，即与它相联系的或然条件所选择的信息项，与另外一个表征的条件或然性相同，即有相同的或然条件信息项，那么这两个表征就具有相同的概念内容。

作为一种心理学理论，CP-CR有很多缺陷。至少而言，如果两个概念在**信念**中与之相联系的信息，存在**任何**一点区别，都会导致两个概念在内容上

206 不同。然而，我们似乎能够在信念中改变概念所相关的信息，却不改变概念

本身。而且,因为概念内容与其所属系统的关系相关,这样一种对概念内容的定义(以或然条件相同定义)意味着,不同人之间就不可能进行概念内容的比较。最后一点,计算按指数增长的或然条件也是难点。

CR 理论主要有三种可供选择的方案:(1)**概念**和**观念**两者能够同时并独立获得概念作用;(2)**概念**的概念作用是基础的,而概念构成观念,所以观念也就继承了概念的概念作用;(3)观念的概念作用是基础的,而概念的概念作用是对其能够构成的所有观念的概念作用中所充当的角色。不同的 CR 理论选择不同的方案。

DCTM

CR 理论应用于 DCTM,结合 LOT 假设,能够得出:表征要么是 LOT 中的条目,要么是 LOT 中由条目构成的复杂表征。因此对于上面的方案(1),即概念和观念都是基础的,LOT 可表述为:

> **(LOT-TC)**
> 系统中 LOT 条目("单词"或者"句子")的内容,是由(至少部分如此)这个条目和系统中 LOT 的其他条目("单词"或者"句子")的关系决定的。[观念和概念都是基础的]

对于方案(2),即概念是基础的,LOT 可表述为:

> **(LOT-C)**
> LOT(思维)中**句子**的内容是由构成句子的"单词"(概念)的 CR,以及单词间内在结构关系所决定的;LOT(思维)中**单词**的内容,是由它与系统中 LOT 的其他单词的关系所决定的。[概念是基础的]

究竟是单词间的哪种关系决定了单词的概念作用,并且以此确定了内容呢?很少有人对此明确说明,就算有所涉及,他们也极少是在谈论同一件事。一种很有影响的观点是把 CR 与"推理"关系联系起来。通常意义上的推理关系与命题的真假值相关,因而单词的 CR 主要是通过句子(观念)来定义,而单词(概念)的 CR 是由句子(观念)衍生一般形式,对于方案(3),即思想是基础的,LOT 可表述为:

> **(LOT-T)**
> LOT(概念)中"单词"的内容,是由其在 LOT(观念)中"句子"中扮演的角色所决定的;LOT(思想)中"句子"的内容由句子在系统中的推理作用决定。[观念是基础的]

207

那么,究竟是哪种"推理"起到了这种作用?在这一点上,不同的人同样也有不同的看法(总是重复同样的说法,是不是有些单调?)。总的来说,推理可能包括:(a)仅指(有效的)演绎推理[2];(b)(有效的)演绎推理和(有说服力的)归纳推理;(c)上面(b)中的两种推理以及人的"决策"。

布洛克对观点(b)提供了一个例证,"句子的概念作用的一个重要组成部分是命题内容,或者说,需要以某种方式符合**演绎**和**归纳**推理"(Block, 1986:628)。布洛克同样对观点(c)作了例证,"概念作用只能从间接推理、**演绎推理**或者**归纳推理**、**决策**等诸如此类的因果关系中提炼出来"(1986)。然而,根据前面(见第7章)对于数字表征的讨论,**有效演绎推理**才是获得推理内容的手段,因为只有演绎推理才能揭示出作为概念组成部分的信息,而不只是这个概念所关涉的那些间接的、偶然的信息。如果世界真的是以"作用(R)"的方式进行表征的,那么只通过有效演绎推理就能揭示出概念所蕴涵的真实信息。我们把这些观点进行整合:表征是处于LOT中的,观念是基础的,演绎推理是概念作用的特征,可以得到DCTM概念作用的内容理论:

(DCTM-CR)

1. 如果"作用(R)"在LOT中是句子表征,那么它的内容由"作用(R)"所涉及的(有效)推理关系决定:

(P) 由"作用(R)",推出…… ["作用(R)"是前提]

(C) 由……,推出"作用(R)" ["作用(R)"是结论]

2. 由与"作用(R)"相联结的具体推理关系,确定R的具体内容。

3. 如果"作用(R)"在LOT中是小于句子的表征(概念),那么它的内容由其所有参与构成的观念中所分担的角色决定。

(P)和(C)的具体推理关系确定了R的具体内容。根据DCTM支持者的观点,至少是表征内容的某些方面,如与计算相关的方面,是由它的推理作用确定的。谓词演算的连接词(见第7章)是一个较常见的例子,如"&"(=and)。已知"P&Q",可以推出"P",同时也就可以推出"Q"。由已知"P"成立和已知"Q"成立,可推出"P&Q"。据此,CR理论家认为这些推理规则确定了"&"("and")的概念作用:

208

(DCTM-CR: &)

表征"P&Q"的概念作用由以下(有效)演绎推理决定:

由"P&Q",推出"P"

由"P&Q",推出"Q"

由"P"和从"Q",推出"P&Q"

如果内容由概念角色确定，可以得出："&"("and")的内容是由它所关涉的推理关系决定的，那么"&"("and")的内容也就是由(DCTM-CR：&)所决定的。当然，概念作用还需要使这种方案能够概括**所有具有内容的表征**。这就是说，CR需要提供一种图式(DCTM-CR)，它能够适用于系统中所有具有内容的表征。然而，概念作用的研究者们只是满足于重复(DCTM-CR)，这一点令人遗憾，因此我们怀疑这种思路是否可行。

8.6 DCTM与意识(I)

前文对"意识(consciousness)"这个概念已有提及，现在对其进行更多的探讨，并尝试说明它与DCTM的关系。首先需要指出的是，"意识(conscious(ness))"这个词有很多相关的用法，然而其中有些用法与我们目前要讨论的问题无关：

1. 对环境中**某事物的意识**(consciousness-of something)(也称为"觉察到(awareness of)"和"注意到(attending to)")。例如，"她意识到了CD播放器里传来的噪音。"

2. **元意识**(meta-consciousness)(也称为"高阶觉知(higher-order awareness)"，"元觉知(meta-awareness)"，"内部通达(internal access)"和"自觉(attending to)")，即对自身心理状态的觉察。例如，"她意识到/自觉自己非常喜欢寿司(sushi)。"

3. **现象意识**(phenomenal consciousness)(也称为"体验的(experiential)"和"质的(qualitative)"意识)，指拥有的某种体验：菠萝与巧克力的味道，或者玫瑰与山羊的气味。

4 **意识—觉醒**(conscious-awakeness)，如"中午时，她失去了意识，一个小时后醒来"。 209

5. **自我意识**(self-consciousness)(也称为"自我觉知(self-awareness)")，即当我们意识到我们自己就是意识的对象(这种用法与另外一种"自主意识(self-conscious)"不同，例如，"十几岁的年轻人自主意识较强，所以对其他事情不够主动")。

可能还有其他的用法。本章将主要讨论元意识，在下一章将讨论对某事物的意识和现象意识。

作为元认知的意识

有一种观点认为,意识关涉**某人对其心理状态的认识**。当我们能报告我们低阶心理状态时,我们就处于这样的元心理状态中。因为这种心理现象并不总是存在,所以出现时,比较令人注意。在日常生活中,当我们在做习惯性的事情时,就会处于"自动执行"状态,这种类型的意识是缺失的。例如,总是驾车沿相同的路线上班,一般不会意识到自己正在如此。先前我们并没有意识到我们正在做某件事——突然地意识到了,但是却记不起来我们是如何做到这一步的。在科学中也有类似的例子。

裂脑人

"裂脑"的提法,源于一些有趣的科学实例——一些病人的大脑两半球运用手术切离(Gazzaniga and LeDoux,1978)。正常人的视觉系统使信息从眼睛传递到大脑的路径,如图 8.3 所示。

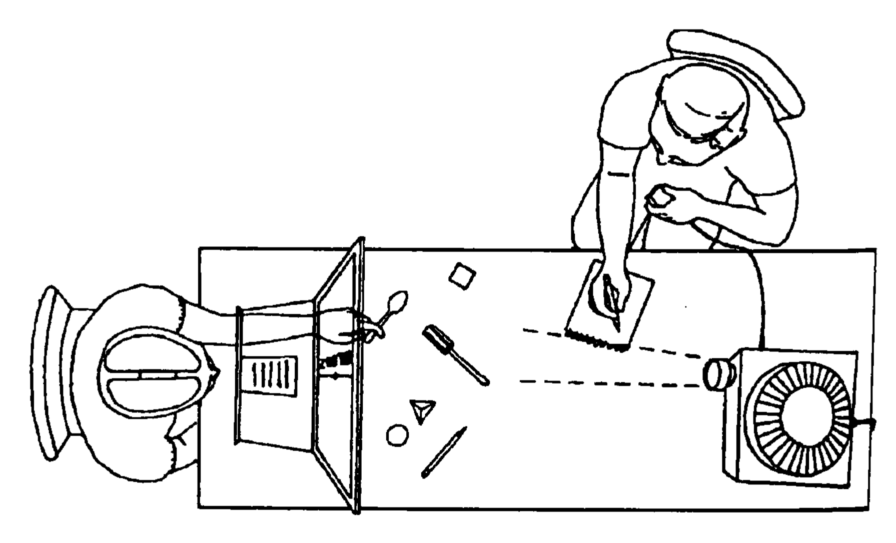
210

左
右
实验中投射的视网膜区域(白色部分)
视觉整体区域
实验中投射的视网膜区域(白色部分)
视神经
视交叉
视束
右侧视觉区域
左侧视觉区域

研究切离脑的视觉,首先要了解它们解剖学上的关系,如图所示。从视觉系统的神经纤维分布来看,呈现给一只眼睛的信息几乎能够对等地投射到两边的脑半球。为了确保信息只投射到其中一个半球,被试需要凝视一个固定点。因为眼球有这样的解剖构造,所以能够使投射到右侧视觉区域的信息只能传递到左脑半球,反之亦然。

图 8.3 视觉偏侧定位(来源:Gazzaniga and LeDoux, 1978: 4, 图 2,图 8.3—图 8.5;经 Kluwer Academic Plenum Publishers 许可复制)

为了使视觉信息对于裂脑人而言仅传到一个脑半球上，实验中物体的图像，需要被投射到右脑半球分别对应的两个眼睛的视网膜区域上。为了做到这一点，使被试凝视一个固定的点上，然后把不同的信息投射到右脑半球对应的两个眼的视网膜区域上（见图 8.4）。

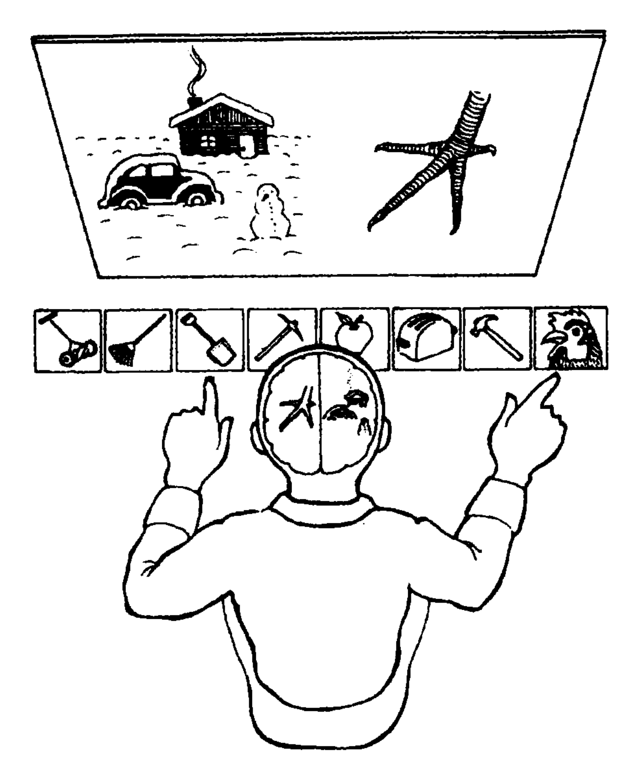


检验视觉偏侧化和实体觉的基本实验设置

图 8.4　裂脑人实验简图（来源：Gazzaniga and LeDoux，1978：5，图 3）

被试的右脑半球能看到东西，但由于语言主要是一种左脑半球的现象（对习惯用右手的人来说），因此不能表达他看到了："切离大脑两半球后，即刻出现的且令人注目的一个现象是，两个半球之间的信息交换被完全中断。因此，视觉的、触觉的、本体感觉的（proprioceptive）、听觉的和嗅觉的信息，能够仅在其中一个脑半球加工处理，并且这些活动即使在另一个脑半球处于觉醒状态中，也能独立进行……只有经大脑左半球处理的信息，裂脑人才能够用语言描述出来，因为左半球通常具有处理自然语言和言语的机制。211 因此，举个例子，如果一个单词（比如'汤匙'）在左侧视觉区闪烁，这个单词只会投射到大脑的右半球……当询问被试时，他会说'我没有看到任何东西'。但是随后被试却能够用左手，从一堆不在视域范围内的物件中，挑选出单词对应的正确对象……另外，如果实验人员问他，'你手中拿的是什么？'，被试会回答'我不知道'。这里再次出现这种现象，主管言语的脑半球说他不知道，这个脑半球没看到图像，也没有获得左手只能传递到右脑的触觉信息。然而显然，右脑半球是知道答案的，因为它对呈现的单词刺激作出了正确的反应"（Gazzaniga and LeDoux，1978：3-5）。另一个著名的实验是，

给被试的右脑半球看一些明显的有关性的图片，左半球没有看任何东西，当问及被试，他们回答“没有任何东西”，但他们会脸红和窃笑。还有其他很多有趣的并包含复杂寓意的实验例子，如图 8.5 所示。



实验运用的方法是同时给被试呈现两个不同的认知任务，每个脑半球都对应着一个独立的任务。左脑半球要求选择鸡爪对应的图片，右脑半球要求选择雪景对应的图片。两个脑半球都作出相应的反应之后，左脑半球负责解释这些选择的原因。

图 8.5　鸡一雪实验（来源：Gazzaniga and LeDoux，1978：149，图 42）

212　这里是部分实验报告：“下雪的景象呈现给被试右脑半球，鸡爪的图像呈现给左脑半球。[裂脑病人被试] P. S. 迅速地作出了正确反应：用右手从右侧的四张图片中选择了一张鸡的图片，用左手也从相应的图片中选择了一张铁铲的卡片。接着询问被试‘你看到了什么？’，被试回答‘我看到了一只鸡爪，我选了鸡，你必须用铁铲来清理鸡棚’。实验重复多次，被试都作出了这种反应。左脑半球能够轻易而准确地确认他为什么选择这个答案，接下来被试眼睛都不眨一下，就把右脑半球的反应整合到左脑半球的框架中。实验人员能够确切地知道被试的右脑半球为什么选择铁铲，但对于被试的左脑半球来说，他对他也选择了铁铲却只是猜测。然而，左脑半球并没有用

猜测，而是用确定事实的口吻回答他为什么同时选择这两张图片”(Gazzaniga and LeDoux, 1978:148-9)。这个实验说明，左脑半球似乎具有元意识，而右脑半球只具有对某事的觉知，至少在这个实验中是这样的。

双耳分听

还有许多心理学实验，在这些实验中，普通被试明显执行了一些**认知**任务，但他们却报告说完全没有意识到。下面是经常被引用的莱克纳和加勒特(Lackner and Garrett, 1972)的实验报告。被试戴上耳机，一只耳朵听到的是模糊的句子，另一只耳朵听到的是具有清晰语境但音量较小使被试难以听到的句子：

右耳　拜访亲戚会无聊(Visiting relatives can be a bore)(模棱两可)

左耳　我讨厌经常来拜访的亲戚与我讨厌出门去拜访亲戚(I hate relatives who visit often vs. I hate traveling to visit relatives)(清晰语境)

要求被试改述右耳听到的句子，另外，如果有被试说听清楚了为左耳播放的句子，那么就让该被试退出实验。当剩下的被试要求重新改述右耳听到的句子时，具有明显的按照并不能听到的句子的语境来解释的偏向。说明被试在没有对句子觉知的情况下，却加工和处理了这些句子。 213

意识与计算机(Ⅰ)

显然是可以制造一台能够对表征进行扫描的机器的，就像一台便携式计算机具有能量管理装置，可以对自我进行扫描以检查屏幕、硬盘等是否开启。从这个意义上说，这样的一台机器可能会“意识到它注意到了”某事，因此只就此而言，这台机器可能是有“意识”的。然而，就像我们将会看到那样，对计算机而言相关意识的问题不止是这些。

8.7　模块化(认知)结构

前面讨论了各种机器组织中的存储和控制，我们把这样的组织称为(机器的)结构。在心智理论中，对组织结构的理解持更宽泛的态度是有益的。这样的理解只是把机器的结构看作是一种特例，它的结构组织可以由人参与。把这种对组织结构的宽泛理解称之为“认知结构”，而且必须把它与具有认知组织结构的系统相联系。

背景:单一结构

如果认为所有的认知(与感知输入和行为输出相对)都有相同的组成部分,那么就可以说认知结构具有一种单一结构。典型的例子是,单一认知模型都可以用两个部分表征心智:在边缘输入(知觉)数据和输出(行为)反应;还有更重要的中枢处理单元。"所有高阶认知功能都能够用这一组基本原则进行解释"(Anderson,1983:2),如图 8.6 所示。

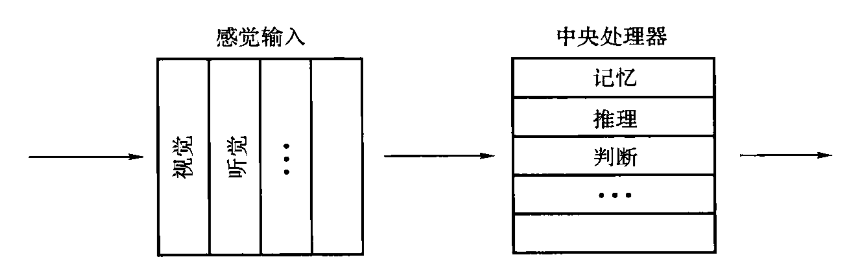


图 8.6 单一结构

214 20 世纪 60 年代,"认知"视角取代了心理学和认知科学中的行为主义,随之而来的一个观点是将心理机能看作心理计算。那个时期,最为人所熟知的计算装置就是具有标准存储程序和寄存器结构的冯·诺伊曼机。逐渐地,心理学中采用计算取向研究认知结构的许多人,就把这种特殊的计算结构等同于认知结构。这种认知计算模型认为,心智的主要部分与冯·诺伊曼机的主要组成部分是一致的(如图 8.7)。

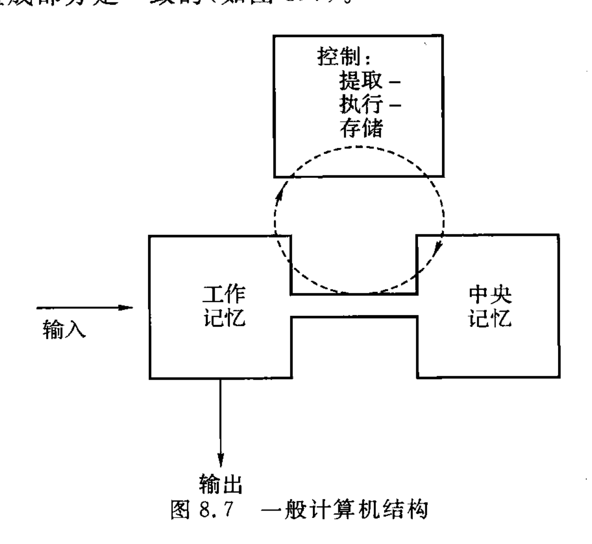


图 8.7 一般计算机结构

冯·诺伊曼机的存储相当于有机体的长时记忆，控制机制根据储存程序指导系统下一步要做什么，输入装置相当于有机体的感知，输出装置相当于有机体行为。

模块化(认知)结构

随后又有另一种认知结构理论被提出，试图解释输入刺激与中枢认知系统的关系。到20世纪90年代，这种理论实际上已经成为认知结构的主流观点："模块性"理论(如图8.8)。 215

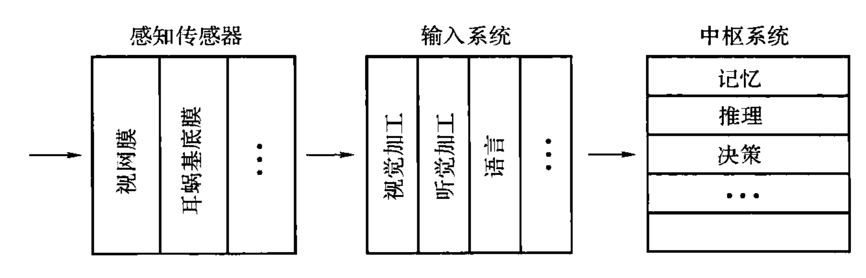


图8.8 模块结构

传统的单一结构，感知输入只与中枢处理相区别。但在模块结构中，又加入了"输入系统"(input systems, ISs)，与感知觉传感器(sensory transducers)[3]和中枢系统(central systems, CSs)相区别。

输入系统模块

根据福多的观点(Fodor, 1983)，人类进化形成了他称之为"输入系统"(input systems, ISs)的心理机能特异种类，这些心理机能可以归类于心理的"自然种类(natural kind)"。因为这些心理机能都共有一些令科学研究者感兴趣的特性，可以概括为输入系统。ISs特性是用功能"定义"的——即系统能做什么，ISs能做的是了解、追踪环境——通过转译感官接触到的刺激，形成按某种法则排列的信息，用来**表征外界事物的特征和结构**。例如，在视网膜上的光(光子)的亮度和分布，变成电信号模式传递到视神经上，最后变成知觉，形成对外界事物特征和结构的表征(见图8.9)。

216

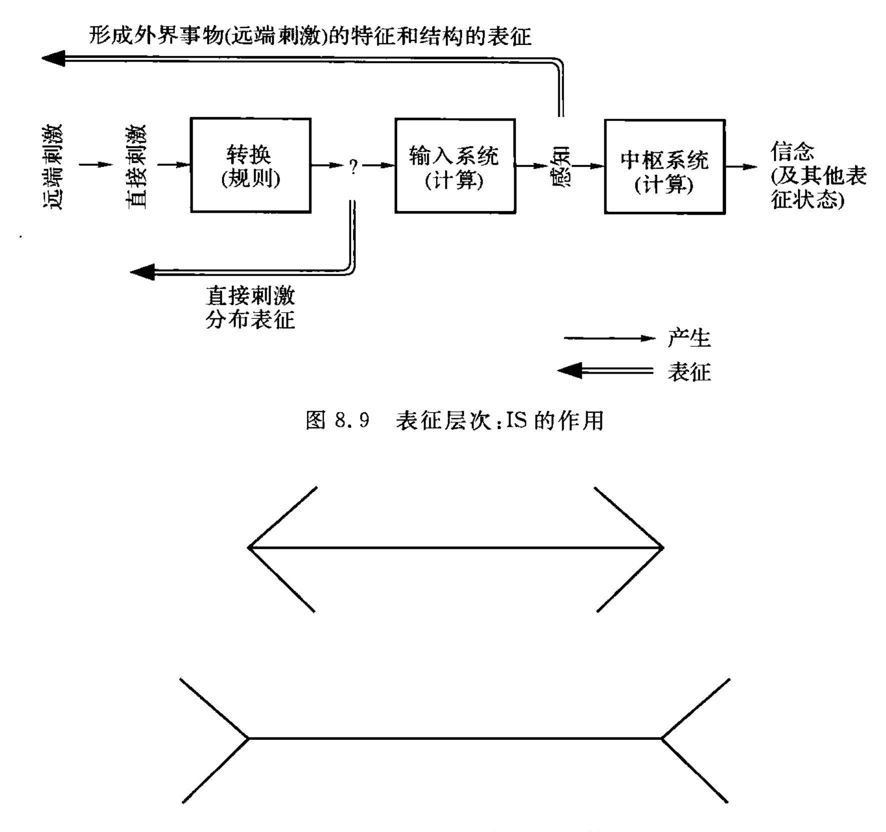


图 8.9　表征层次:IS 的作用

图 8.10　缪勒-莱尔错觉图像

ISs 除了上述的功能特征外,还具有九种在科学上可以研究的非功能性特点,**模块**特征和认知机能**模块化**的程度各有不同。ISs 对(远端)刺激的反应具有**域特异性**(specific domain);还有**强制性**(mandatory),因为 ISs 想要对输入刺激进行操作,就能进行操作;ISs 只能**有限制地通达中枢**,因为如果再想有意识地回溯处理 ISs 的信息是极端困难的;ISs 的运算又是**极快的**(fast),因为它们操作的速度是以百毫秒量级计算的;还具有**信息封装性**(informationally encapsulated),并不能使信息通达其他系统,因为其他系统也许正在做它适于自身的加工处理。缪勒-莱尔(Miiller-Lyer)错觉图像提供了一个很好的例证,知道两条线段的长度相同,也并不会影响对这个图片产生的错觉(见图 8.10)。此外,ISs 还具有**浅输出性**(shallow output),它们快速地向中枢系统传递一种可供内省的"表征层次";它们还与**特定的神经结构**(specific neural structures)相联结;部分是因为 ISs 与特定的神经结构相联结,所以 ISs 还具有**特定的损伤模式**(characteristic breakdown

217

patterns)以及在**发展过程中有特定的步骤和顺序**。

另一方面,中枢系统的功能与 ISs 迥异——**它们的工作是作出合理判断**,或者是对当下情形(如,时一空结构)进行了解,或者决定系统将要做什么(如,如何到达研究生院)。中枢系统并具有上述提及的 ISs 的九种特征,或者至少可以说,没有达到 ISs 的那种程度。CSs 之所以与 ISs 有不同的功能特点,是因为 CSs 还与另两种非常不同的非功能特征相联结。首先,CSs 具有福多称之为的“各向同性(isotropic)”,因为确认一项假设有关的事实可以从系统的**任何地方**提取,例如,并没有任何证据说明萤火虫的亮度和麻醉化学基础一定无关。第二,对于 CSs,福多认为还具有“奎因性(Quinean)”,因为对一个特定假设的确定程度受**整个系统**的系统性、合理性和保守性的影响很大。以上这两种特性都是**整体性**的,因此与 ISs 的运行和功能相区别。

模块类型

福多的模块性只是很多种可能性中的一种而已,还有以下几种能够进行区分。

内部与外部

当一个系统 S,它的内部运行独立于其他系统时(当然,它可能会依靠其他系统的输出作为它的**输入**),我们说它具有**外部**模块性(或者说是相对其他系统而言是模块的)。当一个系统 S,本身可以分解为很多独立的次系统,并且这些次系统对于整个系统而言本身又具有外部模块时,我们说这个系统具有**内部**模块性。与外部和内部模块性相对立的是系统间的同一性或者高度交互性。外部、内部、同一和交互之间进行不同的组合是可能的,这些组合在一些著名的研究程序中都有所体现。例如,系统可以同时具有内部和外部模块性,也可以同时具有内部和外部的交互性(联结主义?),还可以外部是模块性的但内部是交互性的(Fodor),也可以内部是模块性的而外部是交互性的。

硬与软

系统的模块性特征(不论是上面提到的哪一种,也不论如何对它进行确切解释),如果能够归因于系统的“硬件”特征,那么就可以说它是**硬**模块,例如特定线路;如果可以归因于系统的“软件”特征,那么就可以说它是**软**模 218

块，例如不相容的表征图式、无法辨识的数据结构、冲突的控制结构或存储管理系统等。

硬模块的例子，有时候可能用软模块也可以解释——也就是说，可以认为所有电路都具有同一性或者至少可以说不是特定的，但通过“编程”（经由经验或者演变）强加给线路，使它具有了模块性。

附录　模块性：高尔与福多

前面已经简要地回顾了两种心智“模块”观点——第3章讨论了高尔，这一章讨论了福多，如何比较他们的观点呢？

1. 高尔和福多对哪些系统（或者用高尔的术语“机能”）是模块的观点有很大不同。第3章中，已经看到高尔对此作出了确切列举，相比之下，福多带有更多的假设成分：“视觉系统中，能成为模块的可能包括颜色感知机制、形状分析机制和解析三维空间关系机制。也可能还包括执行具体任务的‘高阶’系统，如视觉引导躯体运动或者人脸识别；在听觉系统中，能成为模块的可能包括对殊型言语指派语法描述的计算系统，测定声波排列韵律和节奏结构的计算系统，还可能包括声音识别系统”（1983：47）。对他们的不同观点进行分析，会发现有三个明显特征：（1）他们认为能成为模块的系统，几乎没有重叠；（2）福多的模块比高尔更为具体——涉及较少的常识层面，也很难用常识语言进行简单标识；（3）福多的模块（到目前为止）只限于感知系统，而高尔的模块则更像是说明个性特质、习性和才能。

2. 对于心智如何划分为模块，高尔采用的是（福多称之为）“纵向”机能，而福多认为也可以采取“横向”机能。福多这样描述两者的区别：“横向机能是指对交叉内容域的各种操作，在功能上能够区分为不同的认知系统”（1983：13）。因此典型的，因为人能够进行记忆、注意等多种不同的行为，所以横向机能可以划分为记忆、推理、注意和（先前的）判断等机制。而“纵向机能是针对**具体领域**的，**由基因决定**，与**独特的神经结构**相联结，而且……它们都是**自动计算**的”（1983：13）。

3. 高尔和福多的模块都属于硬模块，但高尔同时认为：（1）对每个人来说，神经线路的位置都是一样的；（2）机能的强弱与机能在脑中相应区域的大小成正比。如果颅骨和脑的相称就如“‘手套与手指相称’一样，颅相学当然就很合理”（Fodor，1983：23）。

219

4. 根据上述第二种观点，高尔的方法是要找出人的心理特质与脑组织

大小之间的对应关系(通过颅骨的隆起),并且每种特质都可以在脑中找到相对应的位置,而且对于每个人来说对应的位置都是相同的。而福多的方法则更具有科学的特征(对于任何给定认知功能,都会在某种程度上呈现出ISs的1—9种特征),而且并没有假定所有人的不同机能对应的不同神经硬件都位于脑的同一位置。这些重要区别归纳如下:

	高尔 vs.	福多
个性特质	Y	N
常识	Y	N
组织大小	Y	N
相同位置	Y	N

注释

[1] 这并不是说物理因果关系就揭开了所有谜团,只是说DCTM只允许存在一个谜团,而不是两个。

[2] 想一想什么是有效演绎推理,例如,“P to Q”具有这样的特征:如果P为真,则Q必定为真——也就是说,P和Q总是具有这种恒常关系。

[3] 感知觉传感器的作用是将外部能量转译为一种能够被心智/大脑识别应用的模式,在后面的讨论中将忽略这一点。

【思考题】

从 RTM 到 DCTM

心智表征理论的两个主要论题是什么?

休谟对于思维和心理表征的观点是什么?

它的优点和缺点是什么?

弗雷格/罗素对于思维和心理表征的观点是什么?

它的优点和缺点是什么?

“命题态度”是什么?举例说明。

命题态度的分类方式——“适切方向”是什么?

命题态度的分类方式——体验特征是什么?

RTM和CTM的关系如何?(即如何从RTM得出CTM?) 220

“形式约束”是什么?

空想先生的例子如何使用命题态度说明CTM理论?

CTM的两个主要论题是什么?

B-DCTM的三个主要论题是什么?

DCTM 与 LOT

DCTM 的三个主要论题是什么?

思维语言(LOT)的三个基本的特征是什么?

支持 LOT“表征方式”的论据是什么?

思维的产生式是什么?

支持 LOT 产生式的论据是什么?

思维的系统性是什么?

支持 LOT 思维系统性的论据是什么?

推理的系统性是什么?

支持 LOT 推理系统性的论据是什么?

DCTM 与心一身问题(mind-body problem, M-BP)

M-BP 是什么?

叙述并评价作为解决 M-BP 方案之一的二元交互论。

叙述并评价作为解决 M-BP 方案之一的副现象论。

叙述并评价作为解决 M-BP 方案之一的行为主义。

叙述并评价作为解决 M-BP 方案之一的同一性理论(物理主义、中枢状态唯物论)(提示:存在两个版本,类型和殊型)。

叙述并评价作为解决 M-BP 方案之一的功能主义。

作为解决 M-BP 的一种方案,功能主义和类型物理主义相比具有哪些优点?(提示:多重可实现性)

221 机器(图灵机)功能主义是什么?与(朴素)功能主义相比具有什么优点?

DCTM 与表征内容

心理内容的“正式观点”是什么?

如何把它应用于分析思考“这个人很高”?

外部的(“长距的”)和内部的(“短距的”)功能性角色(“概念性角色”)是什么?

“形式约束”如何限制作用?为什么?

实现概念作用的两个途径是什么?DCTN 支持哪一个?

概念作用如何解释“and”和“or”等?

DCTM 与意识

元意识是什么?

举一个在日常生活中明显没有元意识参与的例子。

举一个在实验室条件下明显没有元意识参与的例子。

人工智能(机器)可能具有元意识吗?

单一认知结构

单一认知理论的两个主要组成部分是什么?

基于三个机器结构的三种单一认知结构是什么?

它们的优点和缺点是什么?

模块化认知结构

模块化认知结构的三个主要组成部分是什么?

感知觉转译器的功能是什么?

输入系统(ISs)的功能是什么——它们能做什么? 222

IS 的域特异性是什么?

IS 的信息封装性是什么?

中枢系统(CSs)的功能是什么——它们能做什么?

中枢系统(CSs)的两个主要特征是什么?

模块性的两种不同类型是什么?

在哪一点上福多的理论符合这种分类?

【推荐读物】

概论

最近有很多与 DCTM 相关的研究著作、论文和章节。最有影响的长篇演讲报告可能是 Block(1990),也可参见 Block and Segal(1998)。Pylyshyn(1984)对专业的研究者比较适合,Pylyshyn(1989)对这本书作了一些总结。Crane(1995)提供了可读性较强的一般讨论。Glymour(1992)的第 13 章,von Eckardt(1993)的第 3 章,Kim(1996)的第 4 章,Jacob(1997)的第 5 章,Rey(1997)的第 8 章,Cooney(2000)的第 V 部分,这些章节都从不同角度介绍和讨论了 DCTM。

从 RTM 到 DCTM

在 Fodor 的许多著作中都提到了心智表征理论,而比较好的介绍是(1981b)的引言:关于认知科学学科状态的介绍;还有第 9 章:认知科学研究策略的方法唯我论。McCulloch(1995)的第 2 章和第 3 章讨论了经验主义(重点讨论洛克而不是休谟)和弗雷格-罗素关于思维的理论。Churchland(1988)的第 3 章和第 4 章试图将"把握"抽象命题讲清楚。有关命题态度和适切方向的更多讨论见 Searle(1979,1983:第 1 章)。

DCTM 与 LOT

Fodor(1975)最早阐明了 LOT 假设,还考察了一些经验论据以及它对认知科学的影响。Fodor(1987),以及 Fodor and Pylyshyn(1988)讨论了支持 LOT 的论据:产生式和系统性。Maloney(1989)中包含了更多的 LOT 议题,且可读性较强。

DCTM 与 M-BP

223　Fodor(1981a)介绍了认知科学中的心一身问题,且可读性较强。最近多种教科书也讨论了心一身问题及试图解决这个问题的主要心智理论。例如,Churchland(1988)的第 2 章到第 5 章,Kim(1996)的第 1 章到第 5 章,Braddon-Mithchell and Jackson(1996),Rey(1997),Goldberg and Pessin(1997)的第 2 章,Armstrong(1999)等。关于随附性的讨论,参见 Kim 的重要论文(1994),以及 Chalmers(1996b)第 2 章和 Kim(1993)。Kim(1996)的第 4 章对机器(图灵机)功能主义进行了很好的介绍。

DCTM 与表征内容

关于概念(功能)作用语义的介绍及进一步的参考文献,可见 Cummins(1989)第 9 章以及 Lepore(1994)。Field(1977)介绍了将概念作用与主观可能性联系起来的观点。Harman(1982,1987)阐明和维护了“长距”概念作用语义的观点。对概念作用语义的进一步热烈的讨论和辩护,参见 Block(1986)。

DCTM 与意识

关于对意识的讨论的更为完整的书单可见本书第 9 章推荐读物。不过,Guzeldere(1997)还有对意识问题的较好的一般性讨论。Dennett(1991)的第 3 章,Chalmers(1995b)以及 Chalmers(1996b)的第 1 章对意识问题在哲学上的争议进行了较好的介绍。Rosenthal(1986,1997),Lycan(1990)提出了从“高阶”和“内部监控”的意识理论角度讨论元意识,Dreske(1993,1995:第 4 章)对之作了批判性讨论。Block(1995)进一步论述了“通达”意识与“现象”意识之间的区别。

单一性与模块化认知结构

单一性

Block and Fodor(1972)特别是第 II 部分和第 III 部分,对图灵机的认知结构提出了质疑,并支持一种更为一般的计算理论。Newell and Simon(1972),以及 Newell(1973),是关于认知的产生式系统结构的经典研究。Marr(1977)有对产生式系统的批判性讨论。更多基于产生式系统的认知研究,可在 Anderson 的 ACT 系统中找到(1983),更为详尽的介绍参见 Anderson(1993),以及 Klahr et al. (1987)。早期对启发式-PS(PS-inspired)的 Soar 结构的介绍,参见 Laird et al. (1987),而 Newell(1990)是讨论其发展历程的专著。Newell et al. (1989)比
224　较了 Anderson 的 ACT 和 Soar 结构。Lehman et al. (1998)对 Soar 作为一种认知结构进行了介绍。

模块

Fodor(1983)提出并证明存在模块化结构;Fodor(1985)对此作了总结(包括

相关评论和 Fodor 的回应)；Fodor(1986，1989)进一步论述了模块。Harnish and Farmer(1984)，Bever(1992)以及 Segal(1996)区分了各种类型的模块。Bever(1992)，Harnish(1995)以及 Atlas(1997)对福多的模块原理，从多个方面提出了批评。Garfield(1987)收集了很多支持与反对模块性的文献。虽然域特异性是输入系统所具有的重要特征，但 Gelman(1994)认为中枢系统也具有域特异性，Karmiloff-Smith(1992)也持类似的观点——见 Fodor(1998)的讨论。

9 心智数字计算理论的评论

9.1 引言:(再谈)图灵测试

225　我们已经看到了对 DCTM 的精确论述,它为由来已久的心—身问题提供了一种较为合理的答案,为科学研究认知能力提供了一种较具影响力和富有成果的框架。还能从中得到什么呢?另外,作为最早提出的心智理论之一,DCTM 正确吗?著名的"图灵测试(Turing test)"引出了系统是否"能够思维(can think)"这一较受人们关注的问题。参读第 8 章有关内容,从"图灵测试"中可以看到一些引人注意的特征:第一,尽管"智能(intelligence)"这个词出现在图灵的文章标题中,但是他却用"思维(thinking)"进行讨论,这就引起了我们关于下列概念之间关系的疑问:智能、思维、认知、心理等,图灵并没有对这些概念作进一步探究。需要注意的是,当人们倾向于说机器能够正确地进行加法运算,可以下象棋(可能甚至具有智能?),但并不会因此而乐观地说机器会思维,能够深思熟虑,或具有某种心理生活了。第二,我们注意到,当模拟游戏中出现不同性别的助手,他们试图帮助或者愚弄询问者(这里省却了具体的帮助或愚弄的方法),使询问者更容易或者更难辨识是否是机器时,游戏就会变得非常复杂。为什么不只是将"打印机"放在人们面前,然后给他们一段合理时间,最后询问者判断两个与之谈话的交流者中,哪一个是计算机?(尽管图灵在这篇论文中第 6 部分的第 4 项对反对意见的回答中,他似乎认可了一种简化版游戏,即要求询问者判断一首十四行诗的作者是人还是机器。)因为那种简化的游戏安排已经被称为"图灵测试",所以我们使用图灵所说的"模拟游戏"指代这种复杂的游戏安排。第

三，要求询问者辨别哪位是女性哪位是男性，并没有让他辨别哪个是计算机哪个是人类。第四，图灵没有明确地说当计算机赢得模拟游戏时，询问者能够得出什么结论：机器能够思维？它擅长模拟思维？或者其他什么？我们究竟能得出什么结论呢？第五，图灵本人之所以介绍这种测试，是把这个测试当作对“机器能思维吗？”这一问题的替换。那么就会很合理地提出质疑：究竟替换了什么？在那篇文章中，他一方面说问这样的问题是并没有意义的，另一方面却用很大篇幅来叙述这种替换，因此对于图灵来说这种替换还是有特殊意义的。但这种替换究竟意味着什么依然并不十分明确。也许，图灵是想用一个（在科学上）能够“经验证明”的，因而（在科学上）是有意义的问题，替换一个（在科学上）无法证实因而（在科学上）也毫无意义的问题。最后一点，图灵测试具有人类中心主义的（或许文化的）偏见，因为它暗示了智能以及通常意义上的思维，都具有与人类行为无法分离的特征。对思维这一概念的理解是否也存在这种偏见？——思维的标准是不是就应该根据它与人类思维的相似程度而确立？ 226

图灵测试真的很合理吗？假设它是思维的充分不必要条件，图灵似乎就是持这种观点，那么是否可能存在某一事物虽然通过了测试但它并没有思维？最有戏剧性的反例很可能就是布洛克（Block，1990）的“自动会话机（conversation jukebox）”。这种机器（外星人为其编程？还是宇宙中的一个意外？这不重要）似乎具有非常可信的对话能力，其中存储了所有可能问题的答案，但有一些特定要求，比如句子长度的限制（比如说 100 个单词），或者对话长度的限制（图灵测试最初设定的 5 分钟？）。因为有了这些限制，它的程序将会是个有限序列——虽然很长，但还是有限度的。当询问者提出问题 1，自动对话机会从所有储存答案中选择一个恰当的回答，譬如答案 1。当询问者提出问题 2，自动对话机会搜索并选择另一个恰当的答案 2，以此类推。按照这种设计，不可能将自动对话机与人类谈话者区分开，因此这种机器将会通过图灵测试。又因为我们并不认为这种机器能够思维或具有智能，所以自动会话机似乎是图灵测试的一个反例——它能通过测试并不足以表明其具有智能。

显然，机器具有相称的言语行为并不足以证明它就具备了思维能力，或许只能是机器从整体上具备了相称的行为才能成为它具有思维的证据。正如众多人指出的那样，思维与“整体环境的适应”相关，即面对各种可能的环境条件，为实现最终目标，能够进行充分合理的计划、推理和决策。但即使机器能做到这一点，也不一定证明它就具有了思维，因为可能存在一台加强版的“自动行为机”能够合理应对所有可能的环境，就如“自动会话机”能合

理应对所有语言情境一样。但人们在直觉上，还是会认为它并不具有智能。可能不得不作出这样的结论：思维与产生行为的原因有关，而与行为本身的模式无关。接下来我们就对这一结论进行探讨。

9.2 对强人工智能的冲击：塞尔与中文屋

227 塞尔(Searle)的“中文屋”可能是最受人们关注的对DCTM产生了极大冲击的一种论点。很多人对此持反对态度，要么不同意其中心论点，要么认为其中存在某些漏洞(我们都应该抱怀疑态度吗?)。塞尔最早在一篇论文(1980:417)中确切地阐述了他的论点：

> **论点**
>
> 1. 人类(和动物)的意向性是大脑特定因果特性的产物……某些特定的脑过程是产生意向性的充分条件。
>
> 2. 这篇论文的主要目的在于提出这种观点：**实体化**(Instantiating)的计算机程序，不可能依靠自身就具备产生意向性的充分条件。

需要注意的是，“意向性”是一个专业术语，不能将“意向(intentional)”混淆为“意图(on purpose)”(塞尔并没有确切定义什么是‘意向性’，只是说具有“关涉性(aboutness)”的众多心理状态。后面将会进一步对此讨论)。这两个论点又可以产生下列推论：

> 3. 对大脑如何产生意向性的解释，并不能根据这种解释，通过实体化计算机程序而使计算机也具备意向性。这是从(1)和(2)在严格逻辑上得到的推论。
>
> 4. 任何能够产生意向性的作用过程，都必须具有与大脑相同的因果效力。
>
> 5. 任何人尝试创造具有真正意向性的人造机器(强人工智能)都不会成功，除非能够复制人脑的因果效力。

然而，在论文的正文部分，塞尔设立的目标是区分“强人工智能”与“弱人工智能”：

> *弱人工智能*：在心智研究中，计算机的主要价值是为我们提供一种非常有力的工具(1980:417)。
>
> *强人工智能*：认为如果能对计算机进行适当的编程，计算机就会具有认知状态，因此这些程序可以用于解释人的认知(1980:417)。

需要注意的是，这里讨论的是“认知状态”，而不是“意向性”。塞尔提出 228
的这个例子，实际上反对的是关于机器智能的一个相当明确的观点。这个观点认为机器能够运行尚克(Shank)的情境文本(story)理解程序，“1. 机器能够真实地**理解**情境，并能对所提问题进行回答；2. 可以用机器和其程序对情境的理解和回答，来**解释**人类理解情境和回答有关问题的能力”(1980：417)。文章有关这部分的核心观点如下：

(T1)

计算状态**并不是**理解情境，也不是产生认知状态乃至具有意向性的**充分条件**；

(T2)

计算加工**并不能解释**人类理解情境，以及认知过程和意向性的能力。

在考察针对尚克的反例之前，需要先考虑这一问题，即如何从理解一种情境归纳出普遍认知状态。塞尔评论说：“这些论证同样适用于威诺格拉德(Winograd)的 SHRDLU……魏曾鲍姆(Weisenbaum)的 ELIZA……，以及图灵机等对人类心理现象的各种机器模拟”(1980：417)。可以认为，塞尔意思是说，他的这种论证适用于任何“计算”模拟。可以发现，塞尔的论证中有三个潜在的反对目标。第一个也是最直接的对象，是最初尚克提出的情境“理解”程序——这个程序真的能够让机器理解情境吗？第二个也就是次要反对对象是，存在某些程序，能够使机器具有一般意向状态(意向状态是一种包含理解(understanding)和**关涉**(about)某物的心理状态)。第三个也是最不直接反对的对象，是存在某些能够使机器具有一般“心理状态”(有无意向性)的程序。塞尔似乎是想说明，首先，这种反对尚克情境理解程序的论证，能够适用于所有试图用计算机模拟或者解释人类理解情境的能力；其次，他的反对论证可以进一步引申，反对存在某些程序，能够使机器具有一般意向状态，或者所有认知或心理现象。

方法原则

塞尔进行他的反证之前，提出了一项检验心智理论的原则：

(P of I)

检验心智理论是否正确的方法首先要问一问，如果自己的心智，事实上就是按照这一心智理论所说的原则运行，将会出现什么样的情况。(1980：417)

229 显然，该观点认为如果有人提出一种认知功能理论，检验此理论是否正确可以通过自问：如果该理论正确将会怎样？我们把这一方法称为“内省原则”。那么我们将如何理解它呢？想一下句子理解过程的例子。一种句子理解理论可能会包含话语识别的过程，涉及语言起始时间的测算（见 Lisker and Abramson，1964），或者通达心理词汇的过程（见 Forster，1978），以及在所有理解句子的过程中，都有激活与语境不相关联意义的过程。这些理论，因为（1）并不能运用内省原则揭示这些过程，或者（2）想象自己的心智是以这样的方式运行的似乎难以接受，那么，它们就是错误的吗？需要注意的是，我们并不接受（P of I）用“脑”替代“心灵”。那么，我们为什么要接受这一内省原则呢？除非需要假定所有的认知都可以通过内省来揭示（后面还会回到这一问题的讨论）。

塞尔的反证

塞尔想象自己被锁在一个房间里，屋里有一些汉语稿件。因为塞尔不懂中文，所以他也就不知道这些稿件可以分为四组不同的句子。第一组是一篇故事，第二组是针对故事进行解读的文稿，第三组是对故事提出的一些问题，第四组是针对问题所作的回答。最后提供给塞尔一些英文规则，即“程序”，运用这些规则他可以将问题和答案分别对应联系起来。但并不确定房间内的其他懂汉语的人在回答问题的过程中，是否（或如何）参考了对故事解读的文稿。如果没有，中文屋论证就与尚克的程序不相对应，因此我们将假设，中文屋内的其他人在回答问题时参考了解读文稿。当屋外的人用英文向塞尔提出第三组稿件包含的问题时，塞尔根据规则找到该问题在第四组稿件中对应的答案，并将此中文答案再传给屋外的人。通过与屋内懂汉语的人比较，塞尔的行为表现与懂汉语的人并没有什么重要不同，他使用汉语对问题的回答与他用英文一样可靠，由此可以判断他是懂汉语的。塞尔接着总结道：“对我而言，汉语的句子与英语并不相同，我只是通过操作并不理解的形式符号找出答案的。就汉语而言，我所做的就像是一台计算机所进行的工作，对一些具体的形式要素执行一系列计算操作。对于懂汉语的人来说，我仅仅是计算机程序的示例。现在，强人工智能的观点认为，编程后的计算机能够理解故事，程序在某种意义上解释了人类的理解过程”（1980：418）。

中文屋说明了什么：塞尔

塞尔因此得出结论：

1. 关于第一个论断，这个例子表明，我明显一点都不清楚那个汉语故事…… 230

2. 关于第二个论断，即程序解释了人的理解过程。我们不难看到，计算机及其程序没有提供理解汉语故事的充分条件，因为计算机和程序只是在那里运行，这中间并不存在什么理解过程。但它们是否为理解汉语故事提供了必要条件或重要帮助呢？……没有任何理由认为它们是必要条件，或者它们对于理解汉语故事作出了重要贡献。(1980:418)

意向性与脑的因果属性：模拟与复制

机器具有怎样的属性，才能使其获得与人类相同的意向状态的能力？对于这一问题，塞尔的回答是，“只有与脑一样，具有相同因果力的事物，才可能具有意向性”(1980:423)。根据塞尔的这一观点，纯形式的模型就不可能产生意向性，因为形式属性本身并不具备因果力，除非机器的程序从一种状态进入下一种状态能够称为因果力。塞尔认为，这种因果力对于产生意向现象是必备的，而因果力并不等同于驱动机器进入下一种状态。但塞尔还是谨慎地承认，存在某些物理或化学过程可能会产生意向效应——是否真能产生，是能够运用经验证实的。因此事实上，与上面的引述正好相反，塞尔只是强调，如果想要复制脑的意向能力(感知、行动、理解、学习等)，只需要复制**足以产生这些意向效应对应的相关因果力**，“如果能精确地复制原因，那么就能复制结果”(1980:422)。由于大脑还具有多种因果力，并不直接与意向现象相关联，例如生命一维持机能(如果这种机能遭到破坏，可以用生命一维持设备来取代)，因此复制人的意向性并不需要复制脑的所有因果力。这里就有一个从经验可实证角度提出的问题：究竟脑的哪些能力，因果相关于意向性？塞尔并没有提供答案，目前也无人知晓。

在文章最后，塞尔讨论了这样一个问题：为什么当研究者们涉及认知时，比如说当不是气象现象时，就会认为**模拟就是复制**呢？——计算机模拟消化功能时，并不能真的消化吸收一块比萨饼。塞尔认为，首先，这些研究者过于看重类比：心智对于大脑就如同程序对于硬件(1980:423)。这种类比在两个方面上是不适用的，我们已经多少给予了提示：它忽略了脑的**因果**

力是产生意向状态和过程的充分条件(并不仅是状态转换的充分条件);还忽略了心理现象的**意向内容**。其次,这些研究者因"信息加工"概念而误入歧途。强人工智能的维护者辩称,既然计算机能进行信息加工,人也能进行信息加工,那么人与计算机就具有同样的信息加工描述层次——相同的算法层次。但塞尔认为,"就人能够进行'信息加工'这一意义而言……编程计算机并不能进行'信息加工'。更确切地说,它所做的只是操作形式符号而已……它的……符号对于计算机而言并没有任何意义"(1980:423)。第三,对图灵测试还存在普遍接受的这一事实,说明强人工智能依然残留着**行为主义**色彩。例如,他们认为只要机器能够通过图灵测试,就足以说明机器可以进行思考,但中文屋论证就旨在反驳这种观点。第四,强人工智能还残留着**二元论**,因为"他们想要通过编制程序,复制和解释人的心智。但他们是无法完成这个计划的,除非心智不但在观念上,还要在经验上独立于脑……这种形式的二元论……坚持……有关心智的特殊心理内容与脑的实际特征并没有什么固有联系"(1980:423-4)。

9.3 中文屋内的数字计算心智

现在来讨论中文屋的论证如何反驳我们已经确定的 DCTM。我们将从 DCTM 的两个主要维度进行评论,这两个主要维度分别是:揭示认知**状态**或**过程**类型(例如信念、推理)的计算关系,以及认知状态或过程的**表征内容**。

从系统角度对中文屋论证的回应

中文屋作为一种反例,认为计算状态和过程并不足以获得认知状态(或过程),如相信、意愿、打算等,但该论证还存在一些严重问题。中文屋的居住者(冯·诺依曼或许称其为"操作元件")并不懂中文。但至少**作为中文屋整体的系统**是懂中文的(你的左半脑不懂英语,但是你懂英语)。

塞尔对此加以反驳,他认为确实能够通过记忆将作为程序的英语规则"内化",但他仍然不懂中文:"我还是不懂中文,同样地,更不用说我脑中的子系统能懂了,因为作为整体的脑没有的东西,它的子系统也一定没有。**如果我不懂,那么我脑中的子系统也不可能会懂,因为子系统是我的大脑的一部分**"(1980)。

但这种反驳依然存在问题。首先,不能肯定因为塞尔不懂汉语,所以他

的某个部分也不懂。或许他不幸是裂脑病人中的一员。脑中的"某个部分"能够做一些事,事实上也确实做了,但病人本身却意识不到他的这部分脑所做的事,除了观察他们自己的行为。这类病人典型的表现会是,否认自己有能力察觉和理解某事,而这些事恰恰是他们的分裂脑能够察觉和理解的。这就表明,从整体推论部分,或者从部分推论整体,这样的一般推理原则是存在谬误的:水分子不是湿的,那么它们也就不是液体,但是它们由部分构成的整体(一杯水)却是液体。群体是由很多人构成的,但是其中任何一个人都不能成为群体。如果塞尔不认为"系统"能够理解汉语,那么他应该提出其他的理由,而不该是这个。其次,塞尔评论,对于中文屋的居住者而言,汉语的"子系统(subsystem)"与英语的"子系统"有很多不同,他说:"英语的子系统知道'hamburger'指的就是汉堡包,而汉语的子系统对于屋内的人而言,仅能按照规则知道一个'歪扭符号'对应着另一个'扭曲符号'"(1980)。但我们也许会提出质疑,屋内居住者在获得知识的同时,为什么非要假定其中文子系统的理解内容,必须需要依靠内省通达给居住者呢,为什么依靠行为不可以?这是否说明塞尔认为,如前面所讲的(他的内省原则),所有心理状态原则上都可以通达意识?第三,因为塞尔中文屋的居住者是人,他自身可能就具有所有呈现给他的信息,对于中文屋而言,"既不能说他只具有人的生物功能,也不能说他只代表了程序,还是同时具备两者,这都不充分"(Rey,1986:173)。似乎信息"编程"的方式就是唯一可能出错的地方了,也许程序的**储存**并非是机器"恰当编程"的必要条件。冯·诺依曼机确实差不多可以说是通过将输入指令和数据结构置于存储部件(因此能够"储存"它们)之中,而为机器进行编程的,但按照塞尔描述强人工智能的特征,并不要求其程序必须在冯·诺依曼机上运行——毫无疑问,程序必须在冯·诺依曼机上运行是典型的支持强人工智能的机器实践者们的想法。在后面我们将会看到,"联结主义"机器的编程与之非常不同。

从机器人角度对中文屋的回应

也许塞尔的观点已经表明,算法关系并不是产生理解或意向性的充分条件——关涉世界中现实的事物这一意义而言。塞尔说道:"意向性可以定义为某种心理状态的特征,通过这些心理状态,人的心智能够指向(direct)或者关涉(about)**世界中的**客体和事物的状态"(1980:424)。但 DCTM 同时包含语义层次和句法层次。根据 DCTM,理解是指**对于表征的一种计算关系**,该表征是一个(也许会很复杂)具有**表征内容**的符号——关涉某事物。因 233

此，对塞尔论证的回应是，给中文屋增加“机器人(robot)”的感应器，如TV摄像机，这样中文屋就具有了关涉外部事物的能力。

塞尔似乎认为，因为理解的表征内容并不是一种算法，所以理解并不是一种计算关系。对于“强人工智能”来说，这么认为也许是正确的，但涉及DCTM时却可能是错误的，因为DCTM依赖于**符号**或者**表征**。塞尔(1980:420)认为，即使强人工智能具备外部世界的因果关系也不能解决他所提出的疑问——外部世界的因果关系并不构成编程。这也许恰好能反对强人工智能(见后文)，但却并不与DCTM相矛盾。这里出现的问题可能部分地是由于术语造成的。如我们前面所注意到的，“程序”这个概念可以理解包含狭义的、形式的、句法的含义，但也可以理解为还包含更为广义的句法－并－语义(syntactic-plus-semantic)的含义。按照这种广义的理解，即包含语义，两列在形式上完全相同(同构的)的指令，如果它们关涉不同的事物，那么也许可能会是两个完全不同的程序——例如棋子与坦克和部队。强人工智能只采用程序的狭义含义：认为“恰当的编程”就是产生意向性的充分条件(及其他)，也就意味着适当的**算法关系**就是产生意向性的充分条件(及其他)。中文屋当然会对此产生怀疑，但那并不是DCTM的立场，它是在更广泛的意义上看待程序与编程的。

最后，塞尔在他文章的某处写道，“假定，在我不知道的情况下，给我的中文符号来自于TV摄像机”(1980:420)，给机器人增加摄像输入特征就好像真的增加了系统对中文的理解，这其实是误解了心智的输入特征。心智系统是以一种非常特殊的方式接受**感觉信息**的——有人提出脑接受的视觉信息可能是“准图像的(quasi-pictorial)”(Kosslyn，1980)。此外，心智系统还与外界环境有因果联系。事实上，福多也主张，“完全有理由认为(确定的事实)，正确输入特征的因果关系包括，(一方面)联结脑与转换机制以及(另一方面)联结脑与远端事物”(1980b:431)。塞尔认为，单独的因果关系并不足以使系统产生意向性——因为，系统“必须能够，例如，对符号和符号所指向的对象之间的因果关系，产生某种**觉察**(awareness)。因此，现在……我们就不得不放弃强人工智能和机器人的回应”(1980:454)。那么，我们可能会想知道：(1)为什么必须要对符号与其所指之间的**因果关系**有觉察呢，只是觉察到其所指不可以吗？(2)觉察是什么——是意识吗？如果是这样，塞尔是在认为(目前还没有人讨论)意向状态必须要具有意识或潜意识吗(见后文“联结原则”)？(3)如果确实如此，那么为什么他会认为(目前还没有人讨论)意识为什么不能解释为计算？我们稍后将回到意识问题的讨论。

234

中文屋与发光屋

最近，塞尔(1991)以下列形式明确重申了他的中文屋论证：

前提 1　计算机程序是形式的(句法)。

前提 2　人的心智具有心理内容(语义)。

前提 3　句法本身既不构成，也不能满足语义。

结论 1　程序既不构成，也不能满足心智。

塞尔随后扩展了他的论证，提出新的原则，并得出另外三个“结论”：

前提 4　大脑能够产生心智。

结论 2　其他任何能够产生心智的系统，都必须具备某些因果力，这些因果力(至少)要等价于脑的(相关)因果力。

结论 3　任何能够产生心理现象的人工智能装置，任何人造脑，都必将复制脑的特定因果力。仅仅依靠运行形式的程序都不可以。(来自结论 1 和 2?)

结论 4　人脑实际产生心智现象的方式，并不只是凭借运行计算程序。(来自结论 1 和前提 4?)

对此，丘奇兰德夫妇(Churchlands，1991)用“发光屋”的类比作了回应：设想一个人尝试验证这样一个假设，即光是一种在黑暗房间内摇动磁铁产生的电磁辐射(摇动磁铁产生光，就如运行程序理解中文一样)。因为这里产生的光不在人类可见光谱范围之内，所以这个人不会看见任何光。因此，他们可能会得出摇动磁铁不可能产生光这一错误结论：

前提 1　电和磁是某种能量。

前提 2　光的重要特征是有照亮的属性，使人明见。

前提 3　摇动磁铁本身产生的能量既不构成，也不满足发光。　235

结论 1　电与磁既不构成，也不满足发光。

正如这个糟糕的物理结论一样，塞尔关于中文屋的结论也是蹩脚的人工智能理论。按照丘奇兰德夫妇的观点，之所以得出那样的错误结论，需要从前提 3 找出问题之所在；摇动磁铁并没有显示出任何关于光的属性——这是需要深入研究的。现在回到中文屋论证：发光屋与中文屋的论证具有同样的形式，发光屋的结论会使人对前提 3 产生疑问，因此中文屋的论证也会在它的前提 3 上令人怀疑。

塞尔回应丘奇兰德夫妇由“发光小屋”提出的反对观点时这样说道：光是由电磁辐射产生的，但符号本身与因果属性并没有任何关联。它们没有内在语义——能够解释为中文、象棋、股票市场等等。他认为，这恰恰是丘奇兰德夫妇观点的问题之所在：句法要么是形式化的（与形式、形状和结构有关），要么不是形式化的。如果句法是形式的，那么它就不具备因果属性，因此类比也就不成立；如果句法不是形式的，那么就一定需要有其他事物来完成因果工作，例如硬件——是物理事物完成的，而不是程序。

但强人工智能宣称，程序就是产生心智的充分条件，因此塞尔所反对的并不是强人工智能。显而易见，塞尔的回应首先针对发光屋这个反类比的有效性，接着又回到他坚持认为“句法”不具备因果力的观点。查尔默斯(Chalmers，1996a：327)对此作了诙谐的模仿：

1. 食谱是句法。
2. 句法不会产生食物的松郁特征。
3. 蛋糕是松郁的。
4. 因此，遵照食谱不足以做出蛋糕。

稍后我们再回到对这个问题的讨论，目前学术界有关此问题的争论还处于僵局之中。

评论

从系统角度对中文屋论证的回应，引起了人们对意识经验在认知和思维中起到什么作用的研究；从人工智能角度对中文屋论证的回应，引起了人们对认知语义（意向性）的研究。我们需要更加深入地探讨这两个问题，首先是意识，然后是内容。

9.4 DCTM与意识(II)

236

> 意识似乎是最后的堡垒，它是一种生理的副现象，具有某种神秘属性，是无法进行测量的主观状态——简而言之，是在对心智的研究中，留给哲学家涉足并热衷讨论的领域。那就让他们自娱自乐地将这个“现象学”的不可捉摸的东西，形成一个还过得去的理论吧。
>
> ——丹尼特(Dennett，1978b)

在前一章中,我们讨论了元意识,以及对于元意识DCTM可能会如何处理。很明显,能够建构这样一种机器,它可以对自己内部的表征和外部输出进行适当的扫描。在这个意义上,机器能够"意识"到自己内部的状态。但这种内部高层(higher-order)状态,还未涉及通常对意识概念的理解中所包含的另外两个方面:第一个我们称之为"对某事物的意识",第二个通常称为"现象"意识。

对某事物的意识

如果系统能够**觉察**(aware of)某事物,有时我们就会说它具有对环境中**某事物的意识**(consciousness of)。也就是说,系统能够捕捉环境中相关的某类信息,或许系统能够使用该信息对环境作出反应。从这个意义上说,TV摄像机或工业机器人都是有意识的。但"意识到某事物",这种对意识的理解似乎太弱,所以并没有特别引起人们的关注。甚至与元意识结合起来成为元认知(meta-cognition),称为一种"有意识的"状态,似乎依然很单薄。想一想彭罗斯(Penrose, 1989:410)提出的例子,一台摄像机对着镜子拍摄它自己,究竟在什么意义上能说这台摄像机有意识呢?但是当把"对某事物的意识"与下面要讲的意识概念的另一方面相结合时,就会成为日常生活中有关意识的一种重要类型了。

现象意识

还有一种观念,在前面对意识的讨论中并未专论,但对于意识问题来说也是非常重要的,即事物看起来、触摸起来、尝起来、闻起来、听起来等如何,或者说,这些状态的**现象学现象**(phenomenology)、**质的特征**(qualitative character)或**经验层面**(experiential aspects),在这些状态中事物**是其所是的样子**。有人这样强调这个概念,"如果某种心理状态具有某种**质性感受**(qualitative feel)——与经验相联结的质性,我们就会说这种心理状态属于意识"(Chalmers,1996b:4)。想一想,白色屏幕上有一个红色圆点图案,注视这个圆点一段时间后移走红色圆点,这时人会在白色屏幕上看见一个同样尺寸的绿色圆点[1]。这个新的"绿色"圆点究竟在哪里?它并不在屏幕上: 237

① 这种错觉现象是由视觉中"颜色负后像"效应引起的。当红色或者蓝色刺激物对视网膜的刺激停止之后,感觉现象并不立即消失,如果此时将视线转向白色背景,会产生在白色背景上看到绿色或者黄色刺激物的错觉。——译者注

屏幕是白的;它也不在人脑内:脑中有的只是灰质。有些人说意识中具有对绿色经验的质性特征。还有其他很多名称命名这种特征,包括"现象感受",以及意识的"主观层面"。获得这些现象的、经验的或质的特征(通常称之为"感受质性(qualia)")的方法——意识有怎样的经验层面——是通过观察情境,各种感受质性要么彼此非常**不同**,要么就**一起缺失**。

质性不同

内格尔(Nagel,1974)提示我们其他物种的感官系统与人类的非常不同,如蝙蝠。蝙蝠脑中有它特殊的神经环路,可以发出高频声波,然后利用回声对反射声波的物体进行定位。通过这种方式,蝙蝠能够精确辨认事物的尺寸、材质和运动方向——对捕食至关重要。但这并不会误导人们认为蝙蝠是在用耳朵"看"。内格尔接着促使我们思考:蝙蝠利用回声定位探测(看? 还是听?)一只老鼠,那么老鼠对于蝙蝠而言可能是什么样子呢? 这很难回答——对人类而言,这很可能是完全不同的,也许是永远也无法想象的蝙蝠独有的体验。与人类一样,蝙蝠也会收集有关老鼠的信息(也许它甚至能够意识到自己的意识),但与人类完全不相同的(或者我们无法想象的)是蝙蝠具有它独特的经验特征——蝙蝠的感受质性。

颜色特质的缺失

> 可以假设有这么一个人,他看不见红色,但他通过学习物理学的颜色理论,能够理解这个命题,即"红色具有最长的波长"。但他还是不能像正常的未受过教育的人那样理解命题,"这就是红色"。
>
> ——罗素(Russell,1918:55)

或者我们还可以想一想杰克逊(Jackson,1986)提出的玛丽的思想实验。玛丽是一名颜色视觉专家,懂得所有有关颜色的科学知识,但是她本人却是个色盲(或者说是在完全黑白的环境中长大)。她后来经历了一场手术,使她看见了颜色(或者说她从黑白的世界中脱离出来)。杰克逊认为,她现在知道一些颜色的——她以前并不知道的——质性特征。或者从更加形而上,而不是认识论角度来看,某事物现在突然具有了一些以前没有的特征——玛丽对颜色的经验。事物突然具有了一些新的特征,肯定丰富了我们的世界。

238

盲视的质性缺失

有一种尤为有趣的病例类型,病人失去了一些感受质性,也就是所谓的

“盲视(blindsight)”。在盲视病人(他们的视觉系统已经损坏)眼前有一些“必选”题目(他们必须要进行回答),尽管他们并没有视觉经验,但却能够作出正确的答案,他们甚至会认为问题太过简单:“受试者被说服参加一种游戏,例如,‘如果你认为网格线(你不能看到)是垂直的,就说‘垂直’,如果认为是水平的,就说‘水平’。即使是猜测,你也必须给出一种回答。利用这种必选的方法,可以证明病人能够在他们的盲区内探测到视觉事件,并通过转动眼球定位视觉事件在空间中的位置……不仅能够很好地(如果还不能说像正常人一样)辨别不同方向的网格线(一系列的平行线),而且还能进行一些简单形状的识别……尽管事实上显示出他们具有辨别能力,但盲视者还是会说他们无法‘看见’”(Weiskrantz,1988:188)。

这些结果已经被引申出可能存在另外一种感觉通道,“这是一件非常有趣的现象……与盲视现象相似的是‘盲触’的例子,某物接触到了病人手部肌肤,病人由于病症原因并不能意识到这个触觉刺激……但是明确告诉他有这种刺激,病人却能正确地指出手部刺激的位置”(同上)。

DCTM 与完整意识

完整意识

我们把通常人们理解的且引人关注的意识概念称为“完整意识(Full-blown consciousness)”。完整意识似乎又可以分为两种类型。

其中一种完整意识涵盖我们前面提到的关于意识的三种概念,包括:(1)**对环境中某事物的觉察**(being aware of something in the environment);(2)这种意识一般会具有某种**感觉质性特征**;(3)(具有感觉质性特征的)这种觉察本身又会成为觉察的对象——它会**觉察到对它的觉察**(aware of its awareness)。这种元觉察(meta-awareness)一般认为是人类能够报告他的意识状态和控制其行为的基础。 239

另一种完整意识却属于非感觉(non-sensory)的心理状态——例如处理数学问题或决定在哪所学校就读。在这一情况下,这些状态中会产生某种看起来像什么(like to be)的一些独特的东西——一种认知类型的质性特征:**认知质性**(cognitive qualia)。因此,我们就得到了两种典型的意识情景:

1. 环境—信息→感觉质性←觉察
2. 认知状态←觉察

前面(第 8 章)提到,命题态度可分为表征内容(命题)和态度。有人提出

认知状态的内容总是处于自然语言中，或对图形意象的编码中。如果这种观点正确，那么思维这些内容就如储存了某种感觉质性。关于态度、推理、决策、欲望甚至相信，似乎确实都具有**某种看起来像什么的东西**。

总之，完整意识是一种心理状态，要么具有感觉质性，要么具有认知质性（获得那些经验，看起来像什么），并且人还能够对其所处的这些质性特征的状态进行觉察。

意识与计算机（Ⅱ）

计算机能够具有完整意识吗？第一，前面已经提到，获得对某事物的意识，系统需要以某种方式与现实世界相联系，提取有关它思维对象的信息。从对机器表征形式的讨论中（见第 7 章）我们了解到，数字表征存在一个重要难题——还没有一种计算结构涉及如何提取环境信息。我们还了解到，这是功能主义程序需要解决的一个难题，宽泛地讲，就是如何构建一种程序，使之能获取外部环境的信息（参见第 8 章中提到的“长距”概念的作用）。在 DCTM 中，是按照计算操作程序对符号进行操作的形式来处理这些问题的。但什么样的结构才可以使**符号能够替指**（stand for）某事物——以及什么样的替指能够使其与环境相联系呢？因此，DCTM 在从环境中提取信息的问题上出现了困境。一方面，从环境中提取信息似乎并不是计算程序的一部分；另一方面，DCTM 需要具有内容的表征，而内容可能是由环境决定的。DCTM 目前还没有成功解决这个难题。

240 第二，使计算机能够意识到它自己的内部状态，似乎只需对其进行恰当编程即可。但使它具有质性（感觉的和认知的）的意识却是个难点。根据内格尔和杰克逊等哲学家提出的蝙蝠和色盲视觉专家的思想实验可以得出这样的结论：经验的意识、质性的和现象的特征并不具有物理特征，因此物理主义（物质主义）是不完整的。我们对这些形而上的问题持开放态度。然而，从认知科学角度对这些问题的探讨，列文（Levine，1983）称之为“解释鸿沟”，而查尔默斯（Chalmers，1996b）称之为“难问题”。

解释鸿沟

这个观点是指，用生理学（物理学的）的术语对神经系统的描述与如何解释意识现象之间存在着解释鸿沟。在这个世界上，似乎所有有关神经活动和荷尔蒙分泌的知识都不能解释由红色产生的视觉经验、菠萝的味道以及玫瑰的芳香。脑的特定活动与具体的意识经验的联系之间，存在着一道鸿沟。除非我们能够消除鸿沟，否则我们就不能只用生理学的概念来解释经验。

难问题

认知神经科学中(相对)容易的问题是,解释某种脑活动支持某种认知功能或与它的关联。脑内的某些区域和功能在一定程度上已经确认涉及不同的认知机能(例如 Churchland and Sejnowski,1992)。但难问题是大脑的**这些**活动事件如何以及为什么与**那些**意识现象发生联系。对于这个难题,目前我们还完全束手无策。

有些学者认为,目前我们还没有用于理解意识如何源于物质的概念,所以我们需要创造一些新的概念。他们认为,以往物理主义的概念在根本上是有所欠缺的。麦克吉恩(McGinn,1991)认为,由于生物属性对于人类概念能力的限制,我们将永远不会获得这些概念。而另外一些人认为,我们必须重新审视建构自然世界观念的基石,使之能够适用于对意识的理解。内格尔(1993)确切地说道:"关键的问题是下面的做法是否可行,即寻找一种用于理解和描述意识特征的方法,这种方法既符合一般神经生理学概念的理论结构,又能揭示意识的本质……真的能够发现或创造一种通约且完善的视角,运用它能够使我们理解主体经验和神经生理如何发生内在联系吗?……应该明确,对意识的理想解释,必须要使意识现象能够符合生物系统特征的必要条件——该解释使我们知道这样的系统能够成功地具有意识"(1993:5-6)。 241

塞尔(Searle,1992)认为,意识也属于自然现象,但它是由脑产生(和实现)的:"心理现象由脑中的神经生理过程产生,本身也属于脑的特征。为了把这种观点与该领域其他的观点相区别,我们将其称为'生物自然主义'。心理事件和过程是生理自然历史的一部分,就如同消化、普通细胞核分裂、生殖细胞繁殖分裂或者酶的分泌"(1992:1)。

查尔默斯(Chalmers,1996b)也将意识看作一个自然现象,但他并不认为这仅仅是一种**物理**现象,"按照自然法则,意识经验源于物理现象,但它本身却不是物理现象"(1996b:161)。"一种合理的意识理论,必须能够结合一些原则将物理现象与现象的意识领域联系起来,而这些原则……本身并不能限定于物理定律"(1996b:164)。"存在一套系统规则,能够确保某种给定的物理结构可以伴随产生某种特定的经验"(1996b:170)。查尔默斯将他的观点称为"自然主义二元论"。现在就取决于未来的理论框架能否填补解释鸿沟,并在一定程度上回答难问题,使其能够得以解决了。值得注意的是,我们并不能解释**所有事情**,甚至物理学,"万物的科学",也得按照世界本来的面貌接受一些基本原则。现在还不清楚意识之谜何时才能得到完满解

决。我们将会追随内格尔、塞尔和查尔默斯(忽略他们之间的不同),想象一旦相关的**自然**事实都弄清楚了,意识的难题也会迎刃而解。

因此第三点,人们对这个议题的讨论,对我们理解计算机与**质性意识**(qualitative consciousness)问题的关系有哪些帮助?机器要获得质性意识,可能需要使用正确的材料。按照塞尔的观点,如果意识是自然世界的一部分,那么它就是由神经系统的某样东西所产生的,如果我们能复制它的原因,自然就能复制它的结果。但塞尔强调,那并不意味着机器必须需要建造出与人类完全一样的神经系统——这是一个开放的经验问题。可能仅仅只要复制一些与产生质性意识相关的神经系统的特征即可,而其他与意识经验无关的特征可以忽略。或者按照查尔默斯的观点,如果我们能够建造这样一种机器,它的组织结构能够满足意识的心理一物理法则(psychophysical laws),那么意识就会出现在这个机器中。

总结

242 在充分及完整的意识中(包括意识到)需要具备三点:首先,与环境恰当的联系;其次,恰当的编程;最后,恰当的硬件(塞尔)或组织(查尔默斯)。在原则上,是有可能建造一种具有完整意识的机器的,但我们很可能会不能解释,为什么系统在本质上,具有了这样或那样的物理结构和组织就具有了意识,除了能够指出这样的系统确实显现了意识。

认知与意识:"联结原则"

假设"完整意识"是具有以上特征的,那么完整意识状态与认知的关系是怎样的呢?现在至少有三种悬而未决的观点:

1. **必要**关系:如果一种状态是认知的,那么它一定就是意识。
2. **无**关系:认知状态可能是意识的,也可能是非意识的。
3. 认知状态**原则上**需要**通达**意识(塞尔:"联结原则")。

第一种观点似乎太强了。想一想所谓"倾向相信(dispositional beliefs)"的例子,也就是说对于人脑中已经具有的某种相信,也会倾向于接受或表达这些相信。例如,"野生斑马没有穿雨衣(Dennett)",是不是在读到这句话尚未形成意识之前,就已经相信这是事实?或者再来看看"固定相信(standing beliefs)"的例子,当我们睡着时,是不会放弃相信我们已经相信的那些事情的(可以想一想关于自己的相信,如你的名字、电话号码等),我们在睡觉时并未意识到它们。如果真是这样,那么这些认知状态就没必要一定是意识。

还有我们常见的“自动驾驶”、盲视、脑分裂以及双重听觉等现象，都表明认知可以没有意识。

第二种观点代表了人工智能、认知科学以及认知心理学的主要立场。但最近受到了来自塞尔(Searle, 1990b, 1992: 第 7 章)的挑战。他提出了(3)“联结原则”，理由如下：

前提

1. 所有的意向(心理表征)状态，例如相信或意愿，不管它们是意识的还是无意识的，总是具有一种“表象形态(aspectual shapes)”——客观地呈现世界的**方式**(见第 8 章)。一个人可能相信天上远处的星星是金星，而不是晨星或暮星，即便它们所指的都是同一颗星星。一个人可能会想喝水，而不是想喝 H_2O，即便水就是 H_2O。

2. 如果这些形态并不能通过内省的方式，对意识的表象形态(呈现方式)进行解释，那么就只能依靠下面这两种途径：行为和神经生理。

3. 这些“形态特征(aspectual features)”是无法完全或彻底地只用行为或神经生理的概念进行解释的。任何概念都不足以对表象形态进行详尽的描述——是什么形成了**水**与 H_2O 这两种形态，又是什么形成了**晨星**与暮星(这里塞尔赞同内格尔(1974)和杰克逊(1986)的观点，前面已有讨论)。

论证

4. 假设存在“深层”无意识的意向(心理表征)状态——例如，“深层”无意识的信念。

5. 那么这些深层无意识的意向状态就会有表象形态——客观地呈现世界的方式。[源自 1]

6. 这些深层无意识的意向状态并不存在其他的什么内容，如果有的话，也只是指它们的神经生理现象和/或对行为产生的作用。[源自 2]

7. 但神经生理和行为并不能够产生表象形态。[源自 3]

8. 因此深层无意识的意向状态没有表象形态。

9. 而所有意向状态都具有表象形态。[源自 1]

10. 因此，深层无意识的意向状态并不存在。谨此作答。

关于这个论证，我们可以提出两个疑问。首先，在第 2 步，塞尔似乎没有考虑我们在第 7 章深入讨论的，计算机中对于事物表征的数据结构还可能存在其他的选择。深层无意识的意向状态既不是“生理的”，也不是行为的，从塞尔的立场来看，也不是经验的——因此塞尔的论证怎么能对此忽略呢？其次，与之相关的是在第 3 步，塞尔使用了内格尔和杰克逊的“感受质性”的

243

结论，辩驳从神经或行为分析表象形态的可能性——客观地呈现世界的方式。但也有很多理论家相信，并不是所有的“表象形态”或呈现的方式都具有一种可区分的质性特征。他们倾向于认为，相信某物是 H_2O，可能并没有一种可区分的质性形态，或者相信金星是晨星，也不具有可区分的质性形态。人们可以将其中任何一种与茄子的味道联系起来，至于它们呈现了什么，并不重要。这个争论与第一个疑问是相关的，因为根据 DCTM，缺失的数据结构可能正是用来表征思维中非质性形态的东西。所以，塞尔还没有对这种可能性作出解释，但他可能已经有了一种看法。我们后面还会回到对这个问题的讨论。

244

9.5 DCTM 与心理内容

在前一章已经了解到 DCTM 对于内容的“正式”立场，即“概念作用(conceptual role, CR)”理论。一般的概念作用理论，可以如此表述：

(G-CR)

表征内容是由该表征在其所处的概念系统中的作用所决定的。

通过进一步揭示表征内容所涉及的性质、内容所起到的作用的性质以及内容所处概念系统的性质，有人可能还会得出不同的 CR 理论。我们前面的讨论，将 CR 确定为在演绎推理中所起到的作用，把它与思维语言(Language of Thought, LOT)假设相结合。事实上，推理首先和基本上是通过思维(LOT 的“句子”)定义的，这就意味着 CR 这个概念(LOT 中“单词”和“短语”)由其参与构成的思维中的 CR，所起到的作用决定。

那么，我们就得出 DCTM 内容理论：

(DCTM-CR)

1. 如果“R”是 LOT 中的一个句子表征(思维)，那么它的内容就取决于 R 参与了(有效的)推理关系这样一个事实：
   (P)从 R 我们可以推论…… [R 被用作前提]
   (C)从……我们可以推论 R [R 被用作结论]
2. 具体的(P)与(C)的推理关系，与“R”相联结，提供了 R 的具体内容。
3. 如果“R”是 LOT 中小于句子结构的表征(概念)，那么它的内容由其在参与构成的所有句子中所分担的作用决定。

现在，我们来看看该理论可能存在的挑战。 245

(1)***“分析性”难题*** 程式化(P)和(C)的确切规则辖域是什么？现在我们改写一下布洛克(1986:628)提出的例子，思考下面这两个推论：

(CA)费利克斯是只**猫**→费利克斯是**动物**。

(CM)费利克斯是只**猫**→费利克斯**抓老鼠**。

这两个推论与我们常用的例子很不同：从P和Q推出P，“P”是“P和Q”的组成部分，而“动物”并不由“猫”组成。而且，“和(and)”的真值规则保证根据“和(and)”进行推理的有效性(真值确定性)。那么，是什么确保了(CA)的推理结果呢？它是否只是一个关于概念**猫**的事实，或它是否只是(仅仅)一个有关猫的真值？把(CA)与(CM)对比：发现正是(CA)决定了“猫”的内容，而不是(CM)。或者可以这么说，唯心论者认为不存在物质的东西，任何事物都是心理的。泛灵论者认为存在物理的事物，但任何事物都具有心灵(唯我论者认为他们自己是唯一的存在！)。当其中一个理论家想到“**那儿有只猫**”或“**有一瓶很好的墨尔乐红葡萄酒**”时，他们的思维会与我们具有相同的内容吗(假设我们不是唯心论者、泛灵论者或唯我论者)？泛灵论者似乎**能够作出**我们**无法得出**的推论(**那是一瓶很好的墨尔乐红葡萄酒→那里存在着心灵**)，而唯心论者**无法作出**我们**能够得出**的推论(**那儿有只猫→那儿有个物体**)。这些推理的不同，会造成内容的不同吗？对于这种疑问，我们称之为“分析性”难题——特定推理是否包含于内容“分析”中的难题。

(2)***“相对性(Relativism)”难题*** 如果(DCTM-CR)是正确的，那么内容与表征系统有关(即LOT)。这就会引起疑问：不同人之间的心理内容如何进行比较？对于一种特定的思维内容，不同的人同意或不同意如何可能？难道他们一定具有相同的推理关系，又如何可能？如果LOT对所有人来说都是相同的，也许就是可能的。但是，如果人们在彼此完全不同的经历中，形成了他们各自的心理生活，那么又如何可能？一种回答(Fodor,1975)是：LOT是内在的，因此所有人共有。这种观点，也许会帮助我们摆脱目前的这个难题，但却向遗传学提出了巨大挑战。

(3)***“整体性(Holism)”难题*** **两个人当且仅当完全分享了彼此的思维内容**，才能说他们的思维内容是相同的。那么，是什么阻止了他们仅根据有限信息推出对方完整的心智系统的内容呢？应该是能够找出一些描述内容层次的心理规则的。

(4)***“合成性(Compositionality)”难题*** 按照前面的描述，思维的构成成分是从它们在思维中所扮演的作用获得它们的内容的。但合成性的观点似乎是想说明，复杂表征的内容取决于它的成分表征的**独立**内容以及各个成 246

分之间的结构关系。该观点是说，有意义的整体是由具有独立意义的成分合成的。那么，DCTM-CR 与合成性观点之间是否存在矛盾？

(5)***“真值(Truth)”难题*** 首先，使那些符号具有了语义，并不只是推理**模式**(inference pattern)的原因，因为我们不得不假设推理有效，所以才保留了**真值**。这意味着“P 和 Q”必然与真值规则相关联，类似于前面(第 7 章)关于谓语演算的例子。但真值规则本身，既不是推理过程，在计算术语中也没有明确说明。内容计算理论的前提，需要非计算理论(真值理论，或更一般的“模式”理论)已经得出的重要的相关结果作为支撑。

概念作用与宽内容

真值的争论引发了人们对指涉和关涉问题的讨论，因为如果一种思维内容是真的，当且仅当这种思维内容能真实表征相关的现实世界。正是关涉或指涉关系决定了与现实世界相关的部分。这些与现实世界中的事物的关系是思维内容的重要构成要素，还是外在于思维内容？

改动一下普特南(1975b)所使用的术语，如果一种思维内容理论并不假定除思维本身之外还存在其他什么东西，我们称之为“**窄**(narrow)”内容理论，否则，称之为“**宽**(wide)”内容理论。举一个不会有争议的例子，采用适当的态度，“**相信**单身不快乐”属于窄内容，因为“单身不快乐”，无论真假人们都可以相信，并不需要任何有关真值或指涉的前提；而“不**知道**(意识到、辨认出等)单身不快乐”则属于宽内容，因为涉及“单身不快乐”的事实真假。

DCTM

能够看出，DCTM(前面的“形式约束”)属于窄内容理论，因为对于机器是否相信“单身不快乐”的问题，涉及是否将“单身不快乐”的表征储存到了相信盒(如，长时记忆)。如同我们所见，DCTM 并不能使机器分辨出“相信单身不快乐”和“知道单身不快乐”有何不同，因为真值并不是 DCTM 中的概念。

247 人的思维内容究竟是怎么样的呢？有理由需要到“脑之外”寻找那样的内容吗？——有理由假定人的思维内容属于宽内容吗？或许也可以这样说，有理由假设思维内容外在于心智系统自身吗？普特南(1975b，1981)提出了一系列有趣的例子，用以证明**至少有一些内容是宽的和外在的**。

个案 1　[丘吉尔蚂蚁线路]

“一只蚂蚁正在一块沙地上爬行。它在爬行时会在沙地上留下一条轨迹。非常巧合，这条轨迹蜿蜒曲折，交错纵横，看起来像温斯顿·丘吉尔(Winston Churchill)的漫画像。这只蚂蚁是按照临摹温斯顿·丘吉尔的画像爬行的吗？绝大多数人会说没有。毕竟，这只蚂蚁从未见过丘吉尔或者他的画像，没有目的描绘丘吉尔。它只是在爬行而已(它的爬行线路是没有任何目的的)，一条被我们‘看作是’丘吉尔画像的线路……这条线路对于‘蚂蚁自身’根本没有任何东西”(1981:1)。

个案 2　[树的图片]

“想象宇宙中有一颗星球，那里也有已经进化了的人类……假设他们从未见过树的样子(也许那里的植被是以霉菌形式存在的)。假设有一天，一张树的图片碰巧从一艘路过的飞船落在这个星球上，但飞船并未与星球上的人类接触[或者“假设那幅从飞船上掉下的‘树的图片’并不是真的树的图片，而是无意中泼溅的油漆”]。想象一下，那里的人类困惑图片的样子，这究竟是什么呢？他们开始了各种猜测：一座房子，一个斗篷，或者是某种动物。但是，假设他们的猜测从未接近真实。对我们来说，这幅图片是树的表征。但对于那些人类来说，图片只是表征了他们未知的一种古怪事物、一种性质和一种功能。假设他们中的某个人在看过图片后，产生了一种心理表象，与我们看过图片之后产生树的心理表象一样。可以肯定，他的心理表象不是**树的表征**，仅仅只是一张神秘图片描绘的奇怪事物的表征(不管它究竟是什么)”(1981:3-4)。

个案 3　[孪生地球]

“假设在银河的某处，有一个星球被我们称为孪生地球。孪生地球和地球很相似。事实上，孪生地球上的人们甚至说**英语**。事实上，除了在我们这个科幻小说后面明确提出的不同之外，读者可以认为孪生地球与地球在其他方面**完全**相同。甚至可以认为**孪生地球上一个人对应着地球上一个人**(Doppelganger)——完全一样的复制——孪生的两个地球……但孪生地球　248
有一个特别的地方，认为‘水’并不是 H_2O，而是另一种不同的液体，其化学分子式长而复杂，我们给它缩减为 XYZ。假设 XYZ，在通常的温度和压力下，无法与水区别开来……孪生地球上的单词‘水’意指 XYZ……地球上单词‘水’意指 H_2O……按孪生地球人的使用意义……我们称之为‘水’的东西

不仅仅是水；而……孪生地球人称之为‘水’的东西，对我们来说也不仅仅是水”(1975b:228-9)。

尽管普特南在这里使用单词‘水’进行他的思想实验，但是他已经涉及了思维内容：生活在地球上的人能够也确实看到水想到的是 H_2O，而孪生地球的人却不会，他看到水想到的是 XYZ(虽然也称之为“水”)。我们使用的单词“水”(我们正使用的单词)所指的东西(H_2O)与孪生地球上的我们说“水”所指并不相同。

结论

“甚至是庞大而复杂的表征系统，包括语言和视觉系统，与它所表征的东西也不会有一种**内在的**、嵌入的、不可思议的联系——这种联系独立于它的起因以及说话人或思维者的意向”(同上，5)。“如果人与某类事物或它能用语言描述的某种事物，如**树**，完全没有任何因果联系，那么他是无法指称这些事物的”(1981:17)。

我们的观点与普特南一致，这类表征的内容一定是，至少部分地是，宽的或外在的。值得注意的是，自然概念，如水，特别与概念作用的分析相矛盾。因为这种指涉与其核心“意义”层面——概念**所指**，并没有什么关系。就像普特南(1988)提到的，古希腊人对于水的观念与现在很不同，不过“water”是对“**hydor**”的非常完美的转译，因为无论是 water 还是 **hydor** 都指称同一事物——H_2O。经过文艺复兴，直到道尔顿(Dalton)(以及现在的原子化学)以后，人们开始认为液体的属性就是水呈现的特征——因此，酒精和水银的流动是因为它们也同时具有水的特征。显然，人们对水的观念不同，水也就在人们认识其他事物中扮演了不同角色，但人们想到“水”时的思维内容却是相同的，都指称 H_2O。这是因为，概念作用对于判定关涉(determining aboutness)实际上是不起作用的。

内容与计算的难题

假设上面关于内容的结论是正确的——假设概念作用还没有穷尽心理表征内容，至少存在某些思维内容是与现实世界中的事物(物体、属性、情境)相联系的(见第 8 章)。前面第 8 章中说过，DCTM 遵循福多提出的“形式约束”，也就是说，计算是根据其操作形式、结构和句法等表征的特征定义的，而不是表征的所指和真值等语义特征。因此，心理表征具有下列作用(见 Devitt，1989，1991，1996)：

249

心理句法(mental syntax):满足“形式约束”,不涉及语义特征。

窄内容(narrow content):“脑”内的语义特征。

宽内容(wide content):涉及与现实世界相联系的语义特征(真值条件)。

如果还继续接受下面的这条原则:

(A)

认知科学支持的心智规律是通过计算机制应用于表征的计算原则。

那么就可以获得三种类型的认知理论,只有前两种潜在地满足计算形式约束:

句法认知理论:心智规则仅仅满足表征的非语义(形式、句法)特征;

窄认知理论:心智规则满足表征的窄(脑中的)语义特征;

宽认知理论:心智规则满足表征的宽语义特征。

任何接受 CTM(具有“形式约束”)的人,如果同意原则(A),一定会认可前两种认知取向之一。根据 DCTM,心理状态和过程不能满足宽内容——宽内容与计算的分离意味着还存在一些潜在难题。

1. 如第 7 章所讲到的,我们期望 DCTM 能够遵循(track)表征间的逻辑关系运行,如蕴含关系。但蕴含关系具有宽语义特征——本质上使用了真值概念。那么,计算过程如何遵循逻辑关系——如何论证才能满足事实呢?

2. 再列举一个很不相同的例子。当某一种行为获得恰当的“认知”解释时,希望 DCTM 也能解释。例如,如果(a)你走到冰箱前拿饮料,因为(b)你想喝些东西,并且(c)你相信冰箱里有东西可以喝,那么,我们就希望 DCTM 能够使用计算的心理学原则构建一种解释——这些原则涉及计算表征的计算关系和计算操作。但如果**冰箱**并不属于任何表征的一部分,又该如何处理呢? 250

双因素理论:解决办法?

我们在第 8 章介绍功能主义时,提到了两个功能主义因素:(1)“短距”功能主义(也称为“概念作用”理论),只涉及脑的内部关系;(2)“长距”功能主义,(同时)涉及脑的外部关系。由于形式约束只承认内部功能作用是形式

和计算的，因此我们忽视了外部功能作用。

有一种观点是重新审视心理内容，首先确定窄心理状态的内容，也就是说，确定内部功能作用（计算作用）的“窄内容”；然后确定宽心理状态的内容，即决定真值条件（心理一世界适切方向）的外部功能作用的“宽内容”。根据这种“双因素”的内容理论，心理状态可以同时具有窄内容和宽内容，或者更准确地说，同时具有窄内容和宽内容分工不同的层次。内容的窄层次是内部的，可以用计算语言描述；内容的宽层次是外部的，能够用非计算关系语言描述。按照这种观点，一种正确的心理内容理论必须能够同时具有这两个因素，计算也只是整个理论的一部分而已。

9.6 反认知主义

> “机器”无法有效定义为“[某类物理对象]的成员”，因为确定某事物是否机器取决于那个事物实际用来做什么，而不仅仅看其成分和结构。
>
> ——明斯基(Minsky，1967:3)

251 在塞尔的中文屋论文里，他说道：“脑当然是台数字计算机。因为任何事物都是数字计算机，所以脑也是”(1980:424)。塞尔开始对普遍 DCTM 进行第二阶段的反驳（认知是一种计算），甚至直指“认知主义”——“认知主义”认为，从某种特别意义上讲，大脑是数字计算机。这种反对观点很特别，虽然脑可能是数字计算机，但思维却可能不是计算（塞尔的观点）。塞尔(1992:第 9 章)一开始就提出“绝对的基础问题，如究竟什么是数字计算机？究竟什么是符号？究竟什么是运算规则？究竟什么是计算过程？究竟在什么样的物理条件下两种系统能运行同一程序？”(1992:205)。因为“对这些基础问题还没有普遍一致的观点”(1992:205)，于是塞尔回到图灵，对图灵机给出了简要的非形式描述，接着又补充道，“这就是计算的标准定义”(1992:206)。然而，在图灵的著作中还没有一种“计算机的标准定义”——这就是塞尔也无法提供定义的原因；但对图灵机、图灵机计算和图灵机计算功能都作了标准定义；还对图灵机和其他类型计算机作了明确定义，以及一些关于它们计算能力的证明（参见第 6 章推荐读物）。例如，具有与图灵机不同构造的机器，在某种意义上与图灵机仍具有某种**弱等价**关系，它能够计算图灵机能计算的所有函数。可是还没有被普遍接受为认知科学使用的“计算机”或“计算”的定义。

还不清楚塞尔如何或为什么认为，计算机是通过在句法上指派0和1定义的（其中一种巴贝奇（Babbage）机器就是采用十进制计算的，与最初ENIAC的十进制计算相同）。既然0和1只是一种被广泛运用在各种符号中（书信、计数、制图）非常方便的编码方式，所以我们应该修正塞尔**符号**指派的观点：计算机在句法上是通过符号指派定义的。计算机需要的不仅仅是符号——它至少还需要记忆、控制系统以及某些加工能力——计算机毕竟要计算；它又不仅仅是算法——计算机还能够**运行表达**算法的程序。因此，我们认为上面提到的这些计算机部分，都需要融入到塞尔提出的计算机的绝对概念中：计算机至少是一种具有记忆和控制系统，能够操作符号的设备。

塞尔接着又提出另一个问题：如果有人启动一台计算机，他会发现什么，"如果你打开你家的计算机，你几乎不可能发现任何0和1，或者甚至一个磁盘"（1992：206）。此外，计算机还可由各种不同的材料组成：齿轮、杠杆、液压装置、硅片、神经元、猫鼠奶酪、鸽子等等。总之，认知主义告诉我们，"可由任何事物建构一个系统，它能够完成脑能完成的工作"（1992：207）。这就是所谓的"多重可实现性"，根据认知主义的观点，就如化油器可以用铜或者是钢铸成，某种特定的程序可以在各种硬件上运行。塞尔提出异议："但两者是有所不同的：产生问题的原因是，化油器和恒温箱是以它们能够产生某种**物理作用**的方式定义的，但没有人会说可以用鸽子造出化油器。计算机是通过在句法上**指派**0和1的方式定义的。多重可实现性是指不同的物质能够产生相同的物理作用，而不是指相同的物理作用由不同物质构成的事实。这里，与物理事实是完全不相关的，除了在指派0和1（即符号）以及符号间的状态转换方面"（1992：207）。 252

在塞尔的这篇文章中，"相关属性"可能指的是对于成为一台（数字）计算机需要具备的属性。因此，这是塞尔坚持计算机与化油器之间存在不对称性的关键——多重可实现性对于计算机而言，并不只是只要复制了相关原因，就复制了相关结果的后果，如化油器。但我们坚持认为，如果考虑得当，计算机会在这方面与化油器没有区别。

在继续探讨之前，我们必须将**设计**的计算机与按照某种设计实现的作为**物理客体**（physical object）的计算机区别开来——即区分计算机的"**类型**"和"**殊型**"。给定某一计算机设计（类型），当且仅当X实现了该设计，X就是它的殊型。然而，殊型的物理属性会使它从一种（符号）状态转换为另一种（符号）状态，成为设计的一部分。设计层次上的"句法"包含了实现层次（instantiation level）的物理的属性和关系。为方便起见，让我们将上面引述

的塞尔对于计算机和计算的概念称为计算机和计算的“句法定义”。“计算机通过在句法上指派 0 和 1 的方式进行定义”是这种观点的关键之所在。

“句法不是内在物理属性”的争论

对于计算机和计算的句法定义，塞尔进一步得出两个结论：

(1)***普遍可实现性***(Universal realizability)　**多重可实现性**似乎与**普遍**253 **可实现性**包含相同的规则——任何事物都可能是一台数字计算机，因为任何事物都具有构成它的句法类属(syntactical ascriptions)。可以用 0 和 1 的语言描述任何事物(1992:207-8)。更准确地说，(i)任何**事物**都能够对它进行描述，因此对事物的描述就是数字计算(1992:208)(任何事物都是计算机)；(ii)任何事物都有某种程序，该事物正在运行这个程序——我背后的墙正在运行 WordStar 程序(一种计算机程序)，因为墙分子的移动模式与 WordStar 程序的形式结构是同构的(1992:208-9)(任何事物都具有某种正在运行的程序)。

(2)***句法与观察者相关***　句法特征的类属总与主体或观察者相关，他能将某些物理现象看作是句法的(1992:208)。

塞尔的这些观点，对认知主义产生了一些严重威胁：“这会是灾难性的，因为我们想知道是否存在某些东西能够说明，脑内部本质上并不是数字计算机；是否存在有关脑的事实，可以说明脑是数字计算机？这个问题并不能得到回答，因为任何事物都是数字计算机，所以脑也是”(1992:208)。“认知主义的支持者并不把普遍可实现性看作一个问题，因为他们**没有看到‘句法’并不是质量或重力等事物属性的命名**这样的深层结果。句法本质上是一种与观察者相关的概念”(1990b:27;1992:209;着重号后加)。“认知主义的观点不需要‘句法’概念也可以阐明。系统的物理状态是一种计算状态，与**指派**物理状态的某种计算作用、功能或解释相关”。即使不通过指派 0 和 1 定义，也会出现同样的问题，因为诸如计算、算法和程序等概念并不是系统固有物理特征的命名。计算状态**无法在物理事物内部找到**，它们只是**指派**给物理事物的(1990b:27;1992:201;着重号后加)。显然，塞尔将主体的观察视为对认知主义的反驳，但还是不清楚该观点究竟要说明什么，接下来我们尝试进行详细阐述。

论证 1[1]

1. 计算机通过在**句法**上**指派** 0 和 1 **定义**。

2. 如果有什么东西被指派了,那它就**与观察者相关**。[定义]
3. 因此,一种(特殊的)事物是计算机的事实与观察者相关。
4. 因此,如果大脑真的是计算机,这一事实也与观察者相关。
5. 当且仅当某种东西不是事物固有的,它才与观察者相关。[定义]
6. (a)因此脑不是**内在的**数字计算机。 254
   (b)因此脑不是数字计算机。

我们注意到,按照这种论证会得到两种不同的结论。根据塞尔的论证,脑不是**内在的**计算机,因为某种事物是否能成为计算机(being a computer)与观察者相关。但如果真是这样,那么我的/你的 PC 机也都不是**内在**计算机。可是我的/你的 PC 机确实就是计算机,所以我们想知道:(问题 1)**为什么**成为**内在**计算机如此重要——对于脑而言,为什么只**成为**计算机还不够呢?(问题 2)一种**内在**计算机**究竟是什么**?

问题 1:对于脑而言,为什么只成为计算机还不够呢?

塞尔:因为**普遍可实现性**——任何事物都是数字计算机,所以再去说脑是计算机就很无意义和无趣。

回应:有什么理由相信任何事物都是计算机?

塞尔:再次回到普遍可实现性(1):可以指派 0 和 1 描述任何事物。

回应:在讨论普遍可实现性时,塞尔给我们提供了一个所有事物都是计算机的例子,即物理现实都**在运行某种程序**(墙正在运行 WordStar 程序)。但是,当认知主义者认为大脑是(数字)计算机时,是说脑有计算的能力——计算机就是能够计算的设备,尽管它不需要总在计算。

塞尔似乎预设了:

A1. 如果说某事物在一定时间内运行了某种特定程序(的片段),那么它就是在(那段时间内)运行了程序。
A2. 任何(在一段时间内)运行程序的事物是在进行计算。
A3. 任何计算的事物就是计算机。

他似乎认为:

1. 墙在一定时间内运行了 WordStar 程序的片段。
2. 因此,墙(在一定时间内)正在运行 WordStar 程序。[A1]
3. 因此,墙正在计算。[A2]
4. 因此,墙是计算机。[A3]
5. 如果墙是计算机,那么任何事物都是计算机。

255 我们需要使塞尔的这个结论更加清晰，才能够认识到他在什么地方出错了——那么究竟是什么错误呢？有如下几种可能：

(1)(A1)塞尔似乎得出“所有物理事物都能够认为是正在运行某种程序(的片段)”的结论(1 和 2)有些草率。即使事物能够与正在运行的某程序在某段时间内一一对应，也不能说这个事物正在运行此程序。运行程序需要一些形式化的反事实条件句(counterfactuals)的真值：如果给定程序如此这样的一种不可能实现的输入，那么它将会产生一种如此那样的计算；也就是说，如果给程序同时输入数字 2 和 4，它就会出现“No”，或在某种程序中不存在 Control F6 的操作，那么如果给它输入“Control F6”，它就出现黑体。

(2)(A2，A3)塞尔可能还高估了计算与计算机之间的联结程度。也许有人会为塞尔提供一种对计算更宽松的定义，但主张要成为计算机却需要更多的条件。如我们之前所讨论的，计算机可能需要包含结构、记忆和控制等部分。计算机能够计算是计算机概念的一部分，而不是它正在计算。

(3)DCTM 认为“心智/脑是一种特殊类别的计算机”，这一表述是开放式的，不仅仅是说心智/脑就是“一台计算机”。并不只是像塞尔所认为的那样，DCTM 宣称心智/脑就是冯·诺依曼机。这表明，即使 WordStar 程序的片段与一定时间内墙分子的运动存在映射，也不足以说明墙就是冯·诺依曼机。

(4)塞尔的立场表明，没有任何事实证明 PC 机就是冯·诺依曼机。为什么？因为还**可以被说成是**另外一种机器。计算机还能被说成是门制器，刀具也可被说成是镇纸。塞尔对 DCTM 提出的**挑战**(不是反驳)，使我们对系统的描述，尤其是系统计算的描述需要采取更好的方式，从这一点上来说是很有意义的。那么，什么样的描述才算是“更好”？这些描述需要更加精确，对机器行为需要更精确的解释和更精确的语言。继续看塞尔提出的例子，直觉上将 PC 机描述为正在运行 WordStar 程序的冯·诺依曼机，对比将墙看作是正在运行 WordStar，是更加精确的解释，以及还可获得对其行为的更加精确的描述。挑战即是如何精确地说明原因。

(5)语义：我们知道，DCTM 与强人工智能之间至少有一点重要的区别——DCTM 认为认知状态具有**语义**和**表征能力**，而强人工智能是关于心理和程序关系的理论。根据塞尔的观点，程序是形式的、句法的、非语义的和非表征的。因此，即使塞尔能够说明，任何事物都能被看作是运行了依据形式制定的程序，但是要说它可以被看作是运行了依据语义制定的程序，就行不通了。在这一点上，有趣的是，塞尔关于可以运行 WordStar 程序的墙的例证，没有明显(现实世界中)的语义争论。

256

问题 2：内在数字计算机究竟是什么？

塞尔：与观察者相关（句法）的属性是非内在的。而诸如质量和重力等属性，不与观察者相关，因此是内在的。

回应：一种特殊的计算机正在运行一种特殊的程序，这并不是与观察者相关的事实。计算机按照它所是的方式被编程，这是一种内在的物理事实，就像计算机具有重量的内在物理事实一样。当然，正如计算机能够具有不同的重量一样，计算机也可以运行不同的程序。这里的“内在”意味着“与观察者无关”。

我们已经对第一个论证的两个结论做了回应。没有理由假设脑不是计算机——内在的或其他的。也没有理由假设，关于大脑是（数字）计算机的真实性的断言是没有价值的。

论证 2

另一个针对认知主义的论证，来自塞尔对认知主义成为“自然科学”前景的一些评论：

1. 计算机通过**在句法上指派** 0 和 1 的方式定义。
2. 因此，一台（特定）计算机的句法与观察者相关。
3. 因此，脑可被看作是数字计算机的观点并没有事实根据。
4. 自然科学中，是发现（discover）而不是指派（assign）事物的属性。
5. 因此，“认知主义”不会成为自然科学。
6. 认知主义是被认为是，而且被理解为，一门自然科学。
7. 因此，被理解为认知主义的认知主义并不存在。

大脑是数字计算机的主张属于“自然科学”的一部分吗？当然，这一定会被认为是一种经验事实的观点，但并不是所有关于经验事实的观点都能够成为自然科学的一部分。然而，还不清楚塞尔是否需要这个进一步的步骤，因为可以设想，如果句法不是“内在的”（它是被指派的，与观察者相关的，等等），那么这就不是经验事实的观点，只是一个判断（decision）而已。 257

“句法不具因果力”的论证

塞尔就因果力的问题，对认知主义作了第二种反驳论证：“根据认知主义，大脑加工机制产生的认知被认为是计算的。认知主义认为通过制定程序，就会获得脑产生认知的原因”（1990b：30；1993：215）。“可是难点在于，

像这样被指派的 0 和 1,并不具备任何因果力,因为除了在观察者的眼中,它们并不存在。不同于运行的媒介,被运行的程序并没有因果力,因为程序不是真实的存在,它没有本体,超越了运行的媒介。从物理角度讲,不存在独立的'程序层次'"(1990b:30;1992:215)。与前面一样,塞尔的这些评论是很有建设性的,但并不清楚这个论证究竟是什么。那么,我们先尝试对该论证进行明确说明。

论证 1

1. 被运行的程序,超越了运行的媒介,因而不具备因果力。[2]
2. 因此,"程序层次"、"句法"本身不具因果力。
3. 因此,在编程的计算机中,程序没有能力产生程序状态——它**实现于**(realize in)物理事物,而不是**产生于**(caused by)物理事物。
4. 在大脑中,意向状态是物理(生物)过程的结果——意向状态**实现于**并**产生于**物理过程。
5. 因此,大脑不是计算机,认知主义是错误的。

问题出现在从第二步到第三步的推论过程中。**程序本身**是否具有因果力并不重要,重要的是**编程的计算机**是否具有因果力。在第 6 章中提到,编程的计算机事实上是具有独特因果力的,因此,塞尔的论证在第三步出错了。塞尔需要得到的结论是,计算机的程序状态实现于,而不是产生于编程,但这是错的。如果第三步有误,那么就不存在编程的计算机与大脑之间的不对称。

论证 2

258 塞尔可能会反对,任何程序状态都不具有内在意向(intrinsically intentional),因此在编程的计算机与大脑之间仍然存在不同:

1. 计算机通过**在句法上指派** 0 和 1 定义。
2. 因此,一台(特定)计算机的句法与观察者相关。
3. 因此,大脑可被看作数字计算机的观点没有事实根据。
4. 因此,数字计算机不具有内在符号。
5. 因此,数字计算机不具有内在意向性。
6. 但是,大脑具有内在意向性。
7. 因此,大脑不是内在数字计算机,认知主义是错误的。[3]

在中文屋的讨论中,我们还没有看到塞尔对于大脑具有内在意向性的

论证,在这个论证中也没有看到。但是,我们对宽内容的讨论说明某些论证是需要的——回顾前面普特南关于指称的"魔力"理论的讨论。

句法存在于物理事物中且必然具有因果力

> 机器——不仅仅是硬件,而且是被编程了的活着的机器(the programmed living machine)——是我们要研究的有机体。
>
> ——纽厄尔和西蒙(Newell and Simon,1976)

我们知道,通过程序的运行,运行的媒介就具有了这样的设置,即如果给定控制材料的规则,物理状态会按照合适的、自然的方式逐个进行转换——通过机电规则(最终是物理规则)执行被设计的算法。塞尔没有意识到,编程运行的媒介会改变这个媒介——使媒介具有新的和特殊的结构,因此获得特殊的因果力和结果。正是基于这一点,塞尔才认为计算机是图灵机的一种设计(因为还没有这样的物理对象),才认为计算普遍能够用图灵机的计算方式进行描述——在磁带上指派 0 和 1 的操作——这些都是某种误导。图灵机磁带中的 0 和 1,是(典型的)对数字、字母的二进制表征。但是,冯·诺依曼机将程序编译为位码(bit code)而产生的 0 和 1,是系统触发器结构的表征。既然 0 和 1 是典型的电位差的变化,所以它们就表征了编程机器的一种特殊的物理事实。因为"句法"是具有**原因性后果**(causal consequences)的**物理结构**,我们不是仅仅对其"指派"——物理结构并不"与观察者相关"。在通常意义上,更确切地说,它内在于编程的计算机就如同结构内在于事物一样(因为塞尔列举了诸如质量和重力等"内在"物理属性,它们在当代物理学中是有**相互关系**的两个概念,塞尔所说的"内在"必然意指"**与观察者无关或独立于观察者**")。 259

9.7 DCTM 硬件与大脑

程序是产生认知、心理、智能等的充分条件,这是"强人工智能"理论的一部分,那么唯一涉及硬件的问题是,需要硬件具备足够的因果结构执行程序。而且,硬件完成这项工作的特征是使用冯·诺依曼机的结构——储存程序和寄存器,在计算机上实现。可能会有争论认为这并不是必需的,获得心智的编程方式可能需要另一种完全不同的结构。

但是对这个问题,有人可能会采用略微不同的进路。他会说,到目前为

止，我们拥有的唯一心智都是基于生物的，而且在生物基础上不同人之间的心智具有事实上的相似性。因此，如果人们对自身如何运转，我们的心智/脑如何工作感兴趣，而不仅仅关注心智可能会如何工作，那么人们可能会觉得，将 DCTM 提升为基于硬件的“**人类心智**”模型是比较恰当的。

> *DCTM 硬件*：主要可还原为 **0 和 1 触发器**、触发**寄存器**和由触发器构成的**逻辑间**，以及构成这些部件的**电路**。
>
> *大脑*：主要可还原为如下几种类别：**神经元，连接神经元**和连接神经元的“**环路**”。

DCTM 硬件与大脑：相似点

还没有太多信息可以支持 DCTM 硬件可作为大脑的模型。例如：

260

1. 两者都有许多基本要素。
2. 两者都执行了广义的“计算”功能。
3. 两者都根据电信号操作。

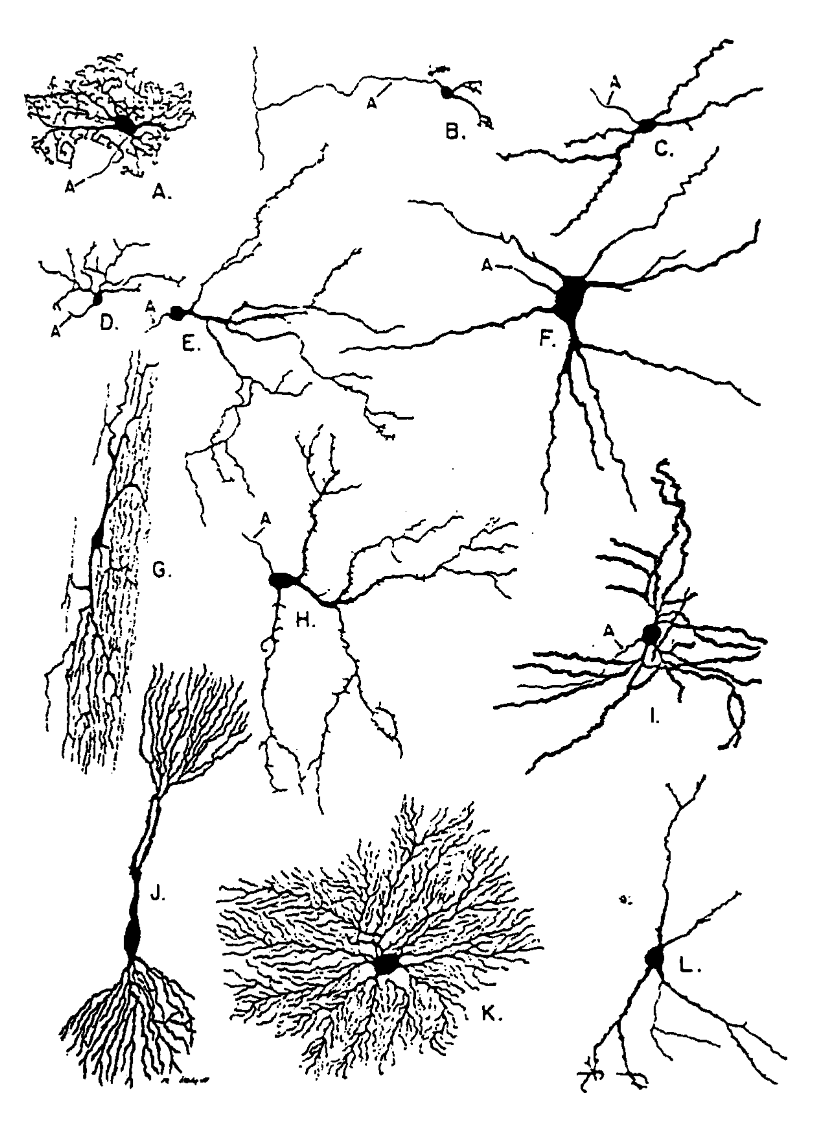
但是这些相似点，同样也可以用来描述微波炉或汽车安全囊。如果我们只是说，存在着一个概括层面，在这个层面上把 DCTM 硬件与大脑联系在了一起，但这个概括层面也适用于微波炉和汽车安全囊，那么就只能认为这种说法没有任何意义了。

DCTM 硬件与大脑：不同点

下面对 DCTM 与大脑的类比进行总结：

构成要素

> 1. 在任何 DCTM 机器中，只有一种相关类型的触发器，但是在脑中却有各种不同的神经元（见图 9.1）。
>
> 2. 触发器只存在开或关两种状态，但神经元具有在生物功能意义上的不同激活频率（回顾第 4 章，蛙眼告诉蛙脑什么）。触发器和神经元还存在从模拟到数字的转换。有一种观点认为，需要很多电子元件才能模拟神经元的主要特征（见 Kent，1981：9，图 1.2）。



几种神经元的树状图。神经结构的解剖构造具有非常吸引人的环路特点。具有鱼鳞状图形的神经元，其神经轴突(A)和树状结构位于中央神经系统之中。A. 细胞核像劣品橄榄球图形的神经元；B. 小脑皮层的颗粒细胞；C. 具有网状结构的细小细胞；D. 脊髓三叉神经核的细小胶质细胞；E. 卵细胞，具有孤束神经核；F. 具有网状结构的大细胞；G. 纺锤形细胞，脊髓胶状物质；H. 脊髓三叉神经核的大细胞；I. 神经元，具有豆状核膜的细胞核；J. 海马皮层阿蒙氏角的双锥形细胞；K. 丘脑神经核细胞；L. 豆状细胞核的苍白球细胞。这些细胞均来自猴，出自高尔基备课材料(承蒙马奎特大学福克斯(Clement Fox)博士允许)。

图 9.1 几种神经元的树状图(来源：Anderson，1995：9，图 1.2；经 MIT Press 许可复制)

3. 神经轴突(与触发器)能够同时对源于不同轴突的神经冲动进行加和(空间加和),还能够对单个轴突的冲动序列进行加和(时间加和),并且每个轴突都是一个独立的信息通道。

4. 与 DCTM 硬件不同,神经元一般情况下能够使用各种神经递质(见第 3 章)。

5. 在一般情况下,神经元会以毫秒的速度工作,而目前计算机芯片是以毫微秒的速度工作。

6. 神经纤维的各种特征(如直径等)会影响神经冲动的传递速度——能够使不同距离的不同神经冲动,在同一时间内到达特定的神经元。而 DCTM 线路则以相同的速度传输信息。

7. 脑内大约有 10^{12} 个神经元,因此,如果每个神经元平均有 10^3 个连接点,那就意味着脑中大约有 10^{15} 个连接点——远远超过 DCTM 硬件。

结构和功能

262

8. 由触发器制成的逻辑门平均有 1 至 4 个连接点,而脑平均是 1 至 1000,最高可达到 1 至 100000。

9. 逻辑门具有固定的几何输入和输出,但神经元却可以生长或去除突触的连接。

10. 逻辑门输出某种节律的频率,对比神经元的输出,是在 0 至 10^2 Hz 内变化。

11. 脑皮层内只有很少的关联神经激活,当然也就没有 DCTM 硬件应用的重复周期。

12. 脑运用注意机制从输入排列中有选择地进行长时记忆,但 DCTM 硬件做不到。

13. 脑中没有与储存所有记忆内容的长时记忆相对应的(解剖)区域,"证据表明,脑在进行记忆时,脑的结构中已储存的信息任何一点被激活,那么新的记忆也便储存于这个结构点之中。脑的记忆可被认为是通过结构分布储存的"(Kent,1981:21)。

14. 脑的控制结构似乎从头到脚遍布全身(见 Kent,1981:13,图 1.5),反映了人从低级结构到高级结构的不同进化阶段。

就这些结构的功能而言,存在一种有趣的共生关系(symbiotic relationship):"……在每个层次上,存在着大量相对独立的子过程,并行进行加工,同时与上下层次交换信息,然后再进行横向交换。因而可以断言,倘若询问脑的任何大范围的功能在何处实现,这种问题是没有意义的。不同功能的方面,实

现于次功能系统的很多不同部分,这些次系统是身体系统的所有主要层次……早期,最简单的脑,显然必须以某种巧妙的方式才能得以生存。在进化的过程中,脑获得了复杂的结构,可以用更加复杂的方式处理同样的基本功能。新的结构不是取代旧的结构并复制其功能,而是仅仅控制旧结构,并将其用作次处理器……不是在需要的时候才将其启动,而是除了在需要时都要抑制其运作。这个系统的美妙之处是,如果高级中心突然受损,原先在正常情况下受抑制的旧的、原始的结构单元的功能,就会解除抑制……对于已受损的低级中心,某些高级结构会对低级中心的大量子加工过程进行再编程,通过这种模拟而掌握低层结构的功能”(Kent,1981:14-17)。

263

上面所引用的只是一种范例,当然它传达的观点是说脑硬件与 DCTM 硬件有很大不同。它们的相似与不同之处当然不是偶然发生的,而是脑与计算机进化的不同。所有哺乳动物的脑基本上都具有相同的神经成分和主要的结构组织(Kent,1981:15,图 1.5)。与普通数字计算机相比,鼠与人类更为相像。所有数字计算机基本上都有相同的触发装置和主要结构(再次参见图 6.5)。与人类相比,IBM 与苹果计算机更为相似。自然选择通过有效成分(神经元)构造了人脑,然后通过使用**如此巨量的神经元**(“你需要百万字节,数万栅极? 没问题,多少兆都行?”)解决即时的**感知**和**行为**问题。商业(非自然的?)选择通过有效但很少应用的(太昂贵了)成分(真空管、半导体)构造了计算机,但是制造它们的速度**非常之快**,最初用于解决**数学和逻辑**问题。因此,脑与计算机的制造方式和它们所擅长的事迥异,这不值得惊奇。

9.8 DCTM 的领域和范围

我们在前面已经看到,DCTM 的充分性在一些前沿问题上受到了挑战,最严厉的挑战可能是针对 DCTM 关于现象意识和宽内容的缄默。因此,我们现在必须面对的问题是:DCTM 的确切适用范围是什么?

心理现象、认知与命题态度

下面列出的是 DCTM 的可能适用范围,但其覆盖程度逐渐降低。首先,如它的名称所表达的,DCTM(注意,M=Mind)是关于所有**心理**现象(状态和过程)的普遍理论。其次,比较保守地讲,DCTM 只是关于**认知**现象的理

论——这些现象直觉上是认知的,涉及心理表征的操作。最后,对于**命题态度**——诸如相信、意愿、渴望等心理状态,DCTM 的范围是最受限制的。对于是否能够或者在何处最后区分出 DCTM 的范围,显然会存在争议。例如,264 对事物的感知是认知的,但不是命题态度。有理由为 DCTM 选择一种更具限制性但其覆盖程度降低的范围吗?有些学者认为应当如此。

是心理或认知现象,而非数字计算现象

某些心理状态和过程,我们直觉上看来是"心理的",不能想当然地认为是数字计算的。下面是初步的、阶段性的考察:

1. **心境**(moods)和**情感**(emotions):根据派利夏恩的观点,"心境和情感几乎肯定要涉及某种非认知因素"(1984:269),所以为了更全面地了解人的心智,我们可能还要再加上某种非数字的计算因素。
2. **睡眠**和**梦**可能无法用数字计算理论进行解释。
3. 人的**创造性**可能不是数字计算过程。
4. 人的**联想能力**可能不是数字计算过程。
5. **认知中的非认知变化**:派利夏恩(1984:266ff)提出,非数字计算的有机体,它的认知存在一些不断发生变化的方式。这些方式涉及"营养的变化、腺和器官(包括脑)的生长和成熟、发生内伤和/或外伤"。福多(1975:200)对"因生理严重损伤而产生的后果"也持同样看法,例如,恶心或感官材料的冲击对认知的影响。这些变化的类型,如同硬件的变化和数字机器的阻断。这些并不属于用纯粹计算算法和表征解释机器行为(尽管很大程度上是对机器行为的普遍解释)的一部分。我们应该允许非认知变化成为"计算"内容的一部分吗?福多和派利夏恩持否定看法。

如果认为上述所列是正确的,那么在直觉上会有一些不是(数字)计算的"心理"现象,所以我们需要将 DCTM 的适用范围,限制在心理生活中的某一部分。从上述所列可以认识到,一般的心理现象能够划分为两类不同的状态和过程,然后根据心理现象的**性质**和它的**历史**又可以划分为不同的次类别,如福多和派利夏恩之前所认识到的,这样就可以将多种心理现象的项目组织到这个列表中。首先,心理状态和过程自身可以划分为两种次类别:经验现象(experiential phenomena)和非经验现象(non-experiential phenomena)。经验现象的特征是具有意识性,通过内省可以通达经验的质性(experiential qualities),以及处于这种状态或过程需要具有某种感官经验

(sensory experiences)。另一方面,非经验现象没有主要的质性特征。显然,许多心理状态和过程是经验和非经验方面的混合,可能甚至是一种谱系 265 (spectrum),譬如说,从"疼痛"递进到"相信"。DCTM所希望的是,这些状态和过程的功能成分具有信息计算内容。其次,心理状态和过程要具有一种时间过程性,要么是认知的,诸如推理、决策等,要么是具有时间过程的非认知事件,这些非认知事件作用于产生认知的物理材料。这种划分如图9.2所示。

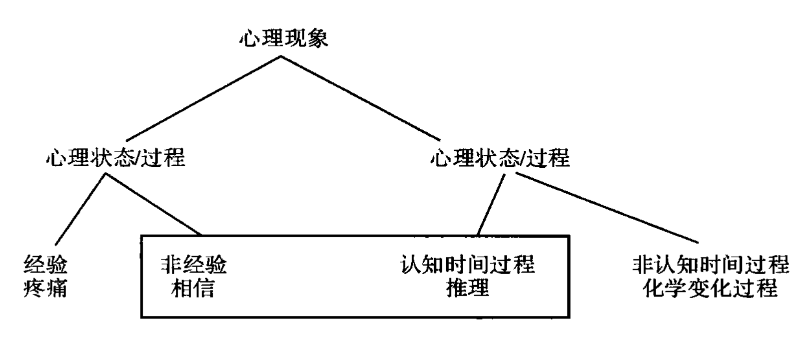


图9.2 DCTM的领域和范围

如果上述划分总体上是正确的,那么DCTM就不会是一种心理现象的普遍理论,而是一种在特定条件下发生某种特定认知现象的理论。如果的确如此,那么可以准确地将DCTM描述为,它是关于认知或因认知而产生的命题态度的数字计算理论。需要注意的是,这也意味着,DCTM并不能解决"身—心"问题,只能够解决"身—认知"问题。科学分析区分出了与前科学稍有不同的领域,这并不值得惊奇。正如福多的一篇相关文章中所提到的:"分子物理维护了直觉上对中等尺寸物体分类为液体和固体的方法。但分子物理也承认了某些物体与液体最为相近,而却在常识上是不能接受的知识,譬如玻璃。可这又怎样呢?"(1987:26)将特殊的认知理论提升为心智的普遍理论,这种倾向可能源于,我们是以思考自己的心理和之所以成为人的那些状态和过程为前提的。与没有认知状态或命题态度的有机体相比,我们更愿意思考不会做梦和睡觉等,缺少心理生活的有机体。因此,古代对"人"的定义包括了"理性动物",而没有包括"能睡觉的动物"或"具有情感的动物"等,这也并不令人意外。传统上,命题态度(相信、意愿、渴望等)构成了我们心理现象概念的核心,因为它们与人的自我身份(self-identity)非常 266 贴近,而其他现象,诸如梦、睡眠、情感、心境和执迷等都被认为是边缘的,当然它们中的任何一个都不能进行合适的科学研究,而只是提供一种方式,使我们有助于思考自身,并认识我们所共有的某些重要东西。

注释

[1] 文章中提出了一种关于“物理事物不具有内在句法的多重可实现论证”：

(1)计算是符号操作——对符号的控制。

(2)符号、符号的同一性和符号的差异性等并非物理定义，

a. 计算状态是多重可实现的，

b. 因此，识别物理事物并不需要识别它的计算状态，

(3)因此，符号的句法不是内在物理的。

但是，这种推论并不成立，因为计算状态是多重可实现的，确定的物理事物并不具有确定的计算状态。思考下面这个类比：母亲能有许多孩子(识别出母亲，并不能确定一个特定的孩子)，因此识别一个特定孩子并不能确定母亲(这个推论显然是错误的)。

[2]从这个意义上讲，如果对于计算机而言是真的，那么对于脑来说，也是真的(复制相关原因，那么就能复制相关结果)。

[3]也许会有这种争论：意向性要么是内在的，要么源于程序；因为意向性并非源于程序(中文屋论证)，所以它是内在的。

【思考题】

图灵测试

描述图灵测试中的“模拟游戏”。

描述图灵测试。

如果机器赢得了模拟游戏，那么图灵认为我们应当得出什么结论？

对图灵测试的主要反对观点是什么？你有何看法？

反强人工智能：中文屋与发光屋

什么是“强”人工智能与“弱”人工智能？

塞尔的中文屋论证想要说明什么？

描述中文屋的**情境**(situation)。

267 塞尔的中文屋**论证**是什么？

塞尔的“脑模拟回应”的内容是什么？

塞尔的回应是什么？

“发光屋”论证是如何反对中文屋论证的？

塞尔的回应是什么？

你认为哪个是正确的？为什么？

DCTM 与中文屋

“系统”的回应是什么？

塞尔的回答是什么？
“机器人”的回应是什么？
塞尔的回答可能是什么？
“模拟”和“复制”的区别是什么？试举例说明。
根据塞尔的观点，为了使机器具有理解能力，产生意向性、认知状态和心理(mentality)，我们应当赋予机器什么？
根据塞尔的观点，为什么研究者们认为强人工智能似乎是有道理的？

意识

什么是对某事物的觉察？
人工智能(机器)能够对某事物产生觉察吗？
什么是现象(质性)的意识？
什么是完整意识？
完整意识的两个典型情况是什么？
什么是“解释鸿沟”？
什么是“难”问题(相对于易解的问题)？
人工智能(机器)能够具有质性意识吗？
认知和意识的三个可能关系是什么？
意识是认知的非必要条件，该观点的理由是什么？
什么是“表象形态(aspectual shapes)”？ 268
表象形态如何与神经生物和行为相联系？
塞尔的“连接原则”是什么？
你赞同连接原则的论证吗？为什么/为什么不？

内容

什么是表征内容的一般概念作用(CR)理论(再次提问)？
DCTM 的内容指称作用理论是什么？
反对 CR 理论的分析性难题是什么？
反对 CR 理论的相对性难题是什么？
反对 CR 理论的整体性难题是什么？
反对 CR 理论的合成性难题是什么？
反对 CR 理论的真值难题是什么？
分别回答什么是“窄”心理状态与“宽”心理状态？
什么是宽内容的丘吉尔蚂蚁路径的例子？请讨论。
什么是宽内容的树图片的例子？请讨论。
什么是宽内容的孪生地球的例子？请讨论。
如果心理内容至少存在部分的宽内容，那么 DCTM 的问题出在哪里？

心理内容的双因素理论是什么？
如何使 DCTM 摆脱宽内容的问题？

反认知主义

什么是“强人工智能”、“弱人工智能”和“认知主义”？
什么是“丘奇-图灵论证”？
什么是“图灵定理”？
塞尔的“计算机的标准概念”是什么？
269 什么是“多重可实现性”？
多重可实现性从何而来？
“普遍可实现性”的两种类型是什么？
普遍可实现性从何而来？
怎样才能成为“内在的”计算机？
关于脑不是内在的数字计算机，塞尔的论证是什么？
认知主义期望获得怎样的认知解释？
这种认知（计算）解释的困难是什么？
认知主义科学研究计划是什么？
我们担忧会出现哪两种后果？
塞尔是如何阐述这些担忧的（提示：蛙眼，语言加工机制）？
关于脑不是数字计算机，塞尔的论证是什么？

DCTM 硬件与大脑

数字计算机和神经硬件（脑）之间有哪些相似之处？
数字计算机和神经硬件（脑）之间**在结构上**有哪些不同之处？
数字计算机和神经硬件（脑）之间**在功能上**（组织和加工）有哪些不同之处？

DCTM 的领域和范围

给出三个我们能称之为“心理的”状态或过程，但是它们不能用 DCTM 进行解释。
我们总结的属于 DCTM 研究范围的心理状态的特征是什么？
我们总结的属于 DCTM 研究范围的心理过程的特征是什么？

【推荐读物】

图灵测试

270 Hofstadter（1981）讨论了图灵测试，此外，Dennett（1985），Block（1990），French（1990），以及 Copeland（1993b）也进行了比较详细的讨论。Feigenbaum and Feldman（1963）第 3 章是关于计算机与智能相关话题的最早经典文集，Luger（1995）是近期出版的关于这个问题的文集。关于智能的“最大适用范围”

的线索,见 Copeland(1993b)第 4 章第 6 节。

中文屋与发光屋

塞尔的中文屋论证已经受到广泛讨论,欲深入地了解该论证,可参见 Searle(1980)中作者对早期有关讨论的评论和回应。把中文屋与认知结构联系起来的讨论参见 Chalmers(1992)。在 Copeland(1993b)第 6 章和第 10 章第 7 节中,同时讨论了中文屋和发光屋,在 Rey(1997)第 10 章第 2 节也有相关讨论。Chalmers(1996b)第 9 章第 3 节讨论了强人工智能,在第 4 节讨论了中文屋。

DCTM 与(现象)意识

在过去的 15 年中,关于意识的研究成果不断涌现。有许多相关著作和论文,还有许多章节和段落的讨论(见第 8 章)。Block(1994)是一部关于意识性质和主要理论的精练而又权威的著作。Güzeldere(1997)内容比较翔实。Goldman(1993b)把意识与认知科学中更宽泛的话题联系起来。Dennett(1991)第 12 章抨击了感受质性和阐述它的思想实验,Lormand(1994)对此作出了回应。

从科学视角探讨的著作(选集)可见 Marcel and Bisiach(1988),Hameroff et al.(1996),以及 Cohen and Schooler(1996)。从哲学视角探讨的有 Block et al.(1997)和 Shear(1997)。

从哲学视角探讨的著作(段落)中,Searle(1992),Flanagan(1992),Chalmers(1996b),以及 Seager(1999)可作为入门读本。如需要了解更多科学视角的观点,可参见 Edelman(1989)或 Crick(1994),然后再关注 Searle(1997)的讨论。

目前在著作的部分章节和段落中进行讨论的有,Maloney(1989)第 7 章,Kim(1996)第 7 章,Braddon-Mithcell and Jackson(1996)第 8 章,以及 Rey(1997)第 11 章。

Levine(1983)介绍了解释鸿沟,Chalmers(1996b)介绍了难问题。这两个问题在 Block et al.(1997)中得到了进一步讨论。

塞尔的连接原则首先出现在 Searle(1990b)中,在 Searle(1992)第 7 章中进行了再次讨论,并被置于一个更大的框架中。在 Davies(1995),以及 Rey(1997)第 9 章第 6 节中,对 Searle(1990)作了评述。关于命题态度的质性感受(被我们称之为"认知的质性感受"),Goldman(1993a)讨论了重要的和被忽视的话题。 271

心理内容和 DCTM

关于概念角色理论介绍性的研究有 Cummings(1989)第 9 章,Lepore(1994)第 4 章。更加详细和深入的讨论可见 Fodor and Lepore(1991,1992)。Putnam(1975b)介绍了宽心理状态,之后在语言哲学中得到了更深入的讨论,见 Pessin and Goldberg(1996)。关于计算和宽内容的问题首先出现在 Stich(1978)和 Fodor(1980a),详细讨论可见 Stich(1983, 1991),Baker(1987)第 3 章,Devitt

(1989,1991,1996:第5章),以及Fodor(1991,1994)。双因素理论(以及宽内容与窄内容)在许多论著中得到了探讨,特别是Kim(1996)第8章,Braddon-Mitchell and Jackson(1996)第12章,以及Rey(1997)第9章。更多对双因素理论进行的辩论,可见McGinn(1982),Block(1986),以及Lepore and Loewer(1986)。Horst(1996)对DCTM作了全面的批判,作者主要关注于符号的因循性(conventionality)。

认知主义

Searle(1992)第9章与Searle(1990b)所阐述的主要观点是相同的。对认知主义的评述可见Chalmers(1995a;1996b:第9章),Copeland(1996a)以及Harnish(1996)。

硬件与大脑

关于该问题的讨论可见Crick and Asanuma(1986),Graubard(1988),特别是Cowan and Sharp,Schwartz,Reeke and Edelman的论文集。Wasserman(1989)附录A包含了一些简要的评论。更多关于大脑和数字硬件"协同进化(co-evolution)"以及对二者进行比较的讨论,参见Kent(1981)。

DCTM领域

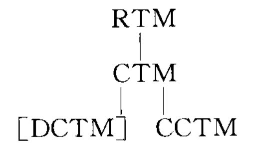
Chalmers(1996b)第1章讨论了称之为"现象"状态和"心理"状态之间的区别,作者所说的"现象"状态和"心理"状态与我们所说的"经验"和"认知"相似。

第三部分

心智的联结计算理论

引　论

第三部分的主要内容是心智的联结计算理论(connectionist computational theory of mind, CCTM)。我们将此理论作为更具普遍意义的心智计算理论(CTM)的两种特殊类别之一(另一种是心智的数字计算理论(DCTM),在第二部分中已有阐述)。此外,在前文中就曾提及,心智计算理论本身又可合适地看作是更古老的心智表征理论(RTM)的一种特殊情形。 275



在第三部分中,首先通过介绍两种在历史上产生重要影响的演示项目——Jets and Sharks 以及 NETtalk(第 10 章),阐述有关联结主义计算的一些基本要点。如果计算机能够按算法操作符号,那么我们就需要关注计算**操作**方面以及计算**符号**方面的问题。这两种网络代表了联结主义模型的两种不同类型:第一种采用了交互激活竞争结构(操作)和定位表征(符号);第二种采用了三层前馈结构(操作)和分布式表征(符号)。之后,在第 11 章中将更深入地介绍联结主义模型的基本建筑砌块(节点和连接),以及它们如何编程,如何进行计算和学习。接下来,综合各种结构和表征方式,概括有关联结主义的一般性看法。最后根据储存、控制和表征的维度对这些结构进行分类。了解了这些有关联结主义计算的主要内容后,我们会对由联结主义计算理论所启示的有关心灵的概念公式化(第 12 章),并作出评论(第 13 章)。

10 联结网络举隅

10.1 引 言

276 在这里，我们将简要回顾两种不同类型的联结网络。一种是在**交互激活竞争网络**中，简单且符合直觉地采用了**定位表征**方式的示例——Jets and Sharks。另一种是 NETtalk，它是一种**三层前馈网络**，因其过于复杂而备受指责。NETtalk 在执行人类任务时，运用的是**分布式表征**，使用反向传播的学习算法。

10.2 Jets & Sharks

麦克莱兰德(McClelland)引入了"Jets & Sharks"[①]网络，为了说明在没有通适规则的系统中表征和储存一般信息的可能性。这种系统能够储存有关对象"样本"及其属性的具体信息，并计算样本与其属性表征之间的传递激活作用，这与前面讨论的语义网络相似(见第 7 章)。当大量样本都被支持时，各属性表征之间的连接就会得到强化。当属性表征之间互相排斥时，则相互竞争。这种解释模型的目的是为了说明，在很多情形下，激活和竞争机制能够：(1)检索某一特殊样本的具体特征；(2)从存储的样本知识中提取有关这类对象的普遍特征；(3)能合理填补缺省值。首先我们来看这个网络的结构和操作，然后介绍它的一些运行情况。

① "Jet"和"Shark"是 1961 年上映的美国著名歌舞剧电影 West Side Story(中译《西城往事》)中两个街头团伙的名称。图 10.1 中出现的 Lance 等人，为剧中两个团伙内的成员。——译者注

277

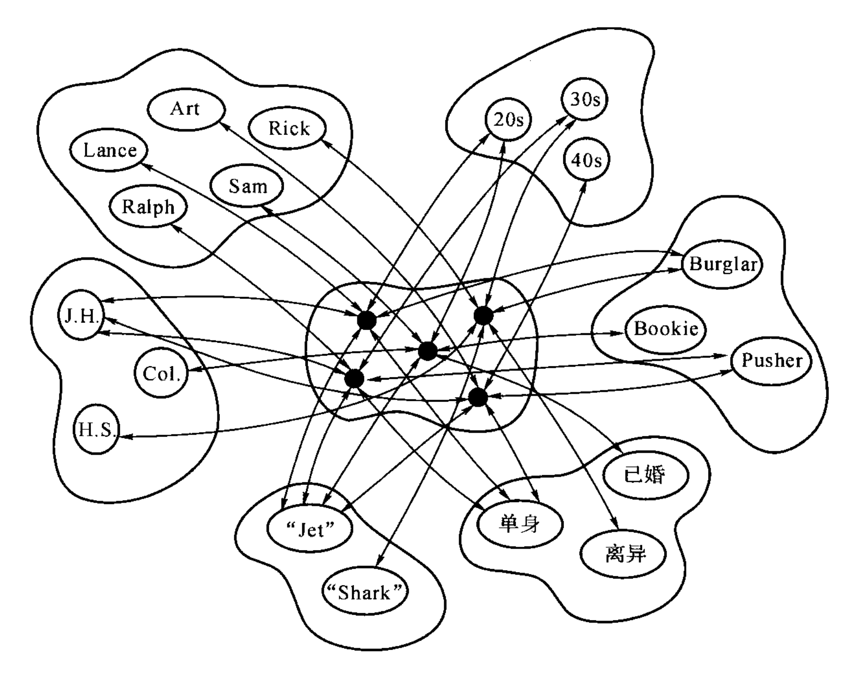


图 10.1　Jets & Sharks 网络模型(来源:McClelland,1981,转引自 Rumelhart and McClelland,1986b,vol. 1:28;经 MIT Press 许可复制)

网　络

如图 10.1 所示网络,包含 7 个组别,共 24 个单元[1]。每个单元:(1)当遇到小于 0 的激活水平时,保持静息状态;(2)**激活遵循的规则是持续不断地加和兴奋与抑制激活值,并连续地输出结果**[2]。其中一个组别是"示例"或称"样本"组(图中的实心单元),表征某一组别里的所有个体成员。每个样本单元与"特征"组中的一个单元,通过相互兴奋而发生连接,如箭头所示。这些连接就表征着网络所具有的有关某一特定个体的特征知识,例如他的名字、教育程度、职业和婚姻状况等等。每一组内的单元相互抑制,这就使每个样本在一个属性组内仅可与其中一种特征发生连接(例如,一个样本不能同时既单身又已婚)。探测针引入之前,每个节点都处于静息状态。当选择一个节点**输入**网络,网络便会使这个节点处于固定激活状态。之后,通过使激活传递到整个网络进行**计算**,使一些单元兴奋,另一些抑制。最终,网络达到平衡态,计算循环结束。那些兴奋的节点为系统输出。现在我们通过网络所执行的下面两个任务,来检验网络行为与人类相比有哪些相似之处。

278

典型特征

当人们询问网络有关“Jets”的信息时，网络会将探测针指向“Jet”，Jet节点便处于固定激活状态。经过200次循环操作后，特征节点出现下面的稳定值：

探测针 Jet
年龄
—1920s：0.663
教育程度
—初中：0.663
婚姻状况
—单身：0.663
职业
—拖运工：0.334
—飞贼：0.334
—赌徒：0.334

这些就是网络所得出的，关于Jets的年龄、教育、婚姻状况以及职业的典型特征。虽然Jet并不具有所有这些特征，但我们仍可以说网络提取了一些Jets的典型特征[3]。

缺省值

网络也能为缺失信息填充缺省值。网络在节点“Lance”和“飞贼”节点279之间受到了损坏。当输入“Lance”时，激活传递约400个循环操作后，网络稳定在下面的数值，意味着它为“Lance”填充了“飞贼”：

姓名
—Lance 0.799
所处团伙
—Jets 0.710
年龄
—1920s 0.667

教育程度

—初中 0.704

婚姻状况

—已婚 0.552

—离异 0.347

职业

—飞贼 0.641

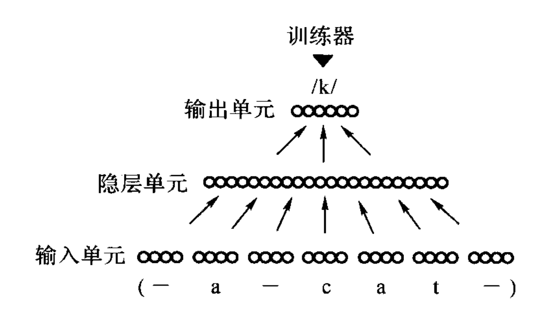
在这个例子中，网络利用那些与 Lance 职业有关的信息猜测她的职业。但通过提高样本节点与样本节点间的抑制值(从 0.03 到 0.05)，用于表示其他个体样本的激活会受到抑制，这样它们的特征节点不会激活，系统也就不会给 Lance 返回缺省的职业信息。这种可变抑制机制会使网络相应地对自身的“反问”发生变化：“我准确地知道关于 Lance 的什么?”与“哪些信息很可能是符合 Lance 的?”

总　结

Jets & Sharks 网络和它的执行过程说明了它具有很多重要的特征，可总结如下：首先，它具有一定的**结构**——包含一定数目的单元与相应的单元间连接；第二，能够实现**表征**——每一单元表征某一个体或其特征。表征采用的是一种定位式(或“定点表征”)；第三，网络具有一定的**计算**方法——通过传递兴奋或抑制，以及竞争进行计算；第四，网络通过具体的联结起所有单元的激活传递规则，实现**编程**；第五，网络不能**学习**，也不能受到**训练**。当后面再介绍其他网络的时候，还会返回来与这种网络的一般特征进行对比。

10.3　NETtalk

NETtalk 是一种能够朗读英语文本的网络，由谢诺沃斯基(Sejnowski)和罗森伯格(Rosenberg)提出。网络训练时，给予连贯文本和配以正确读音的单词字典，能够逐渐对其发音进行调整，最终可达到教学标准。现在我们依据联结主义机器普遍接受的一种划分方法，列出 NETtalk 的几个主要组成部分：结构、表征、计算、编程和学习—训练(如图 10.2)。我们也将依据人类行为对这种模型进行评估。 280



NETtalk 网络结构示意图。英语文本的字母窗格是一排 203 个输入单元。从这些单元获得的信息，由中间 80 个单元的"隐藏"层转换为 26 个输出单元的激活模式。网络的连接由总数为 18629 个权值参数实现(包括每个单元阈值)。

图 10.2 NETtalk 网络结构(来源:Sejnowski and Rosenberg,1987:147,图 1;图 10.2—10.7 经 Complex Systems Publications 许可复制)

静态特征

结构

单元

1. 309
2. 3 层:

   输入层:203 个单元,分为 7 组,每组 29 个单元

281 输出层:26 个单元

   隐藏层:80 个单元

联结

1. 每层单元只连接到下一层单元。
2. 前馈:兴奋(和抑制)从输入单元开始,流向隐层单元,之后传向输出单元。

表征

输入单元

每组 29 个单元:

1. 每个单元对应一个英文字母(26 个单元)
2. 表征标点符号以及词边界单元(3 个单元)

输出单元

26个单元共编码成：

1. 21个音节单位

2. 5个重音以及音节边界点

隐层单元

需要详细分析——将在本章最后部分讨论。

输入和输出表征可以分为两种方式：一种是定位式，另一种是分布式。如果每一个输入单元是定位的，那么输入层也就是**定位式**——每个输入单元表征字母表中的一个字母或者一个标点符号。同样，如果每个输出单元是定位的，那么输出层就为**定位式**——每个输出单元表征一个音节单位、音节分界线，或者重音程度。另一方面，如果输入层是关于单词的表征，那么输入层为**分布式**——表征一个单词需要不止一个输入单元，比表征一个字母需要的输入单元要多。同样，如果输出层是关于整个语音或者音位的表征，那么输出层也为**分布式**——因为一个单词发音（音位）的合成，通常要多于三个音节。因此，表征采用"定位式"还是"分布式"，与网络所要表征的内容相关。通常认为，当说网络属于"分布式"表征时，那么至少它的某些表征是分布式的；当说网络是"定位式"表征时，那么其所有的表征都必须是定位的。

动态特征

计算

1. 输入单元激活值（"输入矢量"）乘以输入单元与隐层单元的联结权值，所得结果传递到隐层单元，作为隐层单元的激活值（"隐层矢量"）。 282

2. 隐层单元激活值乘以与隐层单元与输出单元的联结权值，所得结果传递到输出单元，作为输出单元的激活值（"输出矢量"）。

编程

单元

1. 系统以S形激活进行编程，与神经元的激活大致相似（如图10.3）。

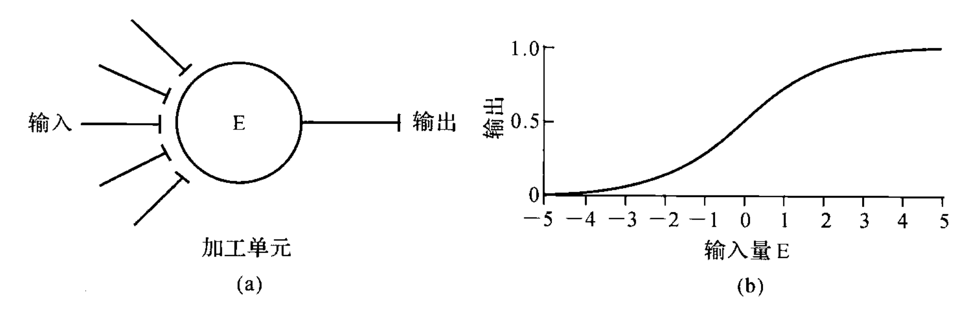


图 10.3 输出函数(来源:Sejnowski and Rosenberg,1987:148,图 2)

2. 阈值可变(后文中将忽略阈值)。

联结

起始联结权值随机分配－0.5 到＋0.5 之间的数值。

学习/训练步骤

283 1. 每个单词呈现给网络后,系统采用“反馈传播”训练。这个过程简要概括如下:

i. 对给定输入(单词)计算网络的输出结果。

ii. 将之与目标输出进行比较,找出误差(目标值与实际输出值的差异)。

iii. 误差反馈。网络把误差首先传递给隐层单元,再到输入单元。

iv. 按某些确定的参数(学习率)调整联结权值以减少误差(差异)。

v. 重复步骤 i—iv,直至目标值和实际输出的偏差可接受。

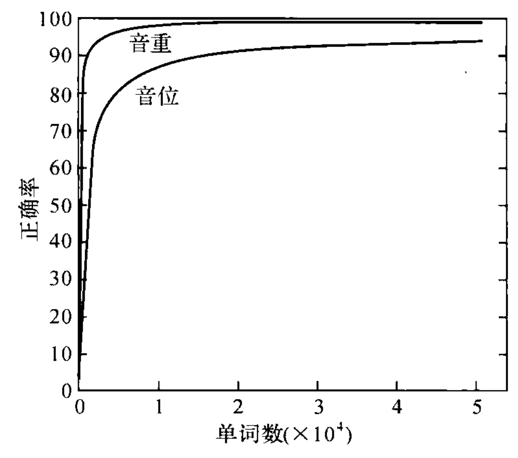


图 10.4 重音、音位学习曲线(来源:Sejnowski and Rosenberg,1987:153,图 4)

2. 呈现给网络的两种不同类型的材料：

i. 包含 1024 个单词的连续文本，来自一年级儿童的非正规独白。

ii. 1000 个常见单词，出自《梅里亚姆-韦伯斯特袖珍词典》，以随机顺序呈现。

3. 材料（文本或单词）按次序移动通过可同时容纳 7 个字母的“窗格”。连续文本包含的字母逐一向前移动，而词典单词则是整个儿地通过窗口。对于连续文本，读取字母的是中间的第四个窗格，其余六个窗格提供上下文——提供所需要的信息。

结果与分析

非正规连续文本运行结果

1. *重音*：重复 5 次 i—iv 步骤后，正确率为 99%。

音位：重复 50 次后为 95%。

网络最早能够区分的是元音和辅音。之后掌握的是单词边界，重复 i—iv步骤 10 次后，就可理解网络的“谈话”了。

2. 综合/总的输出正确率：对于 439 个单词的连续文本，正确率为 78%。

3. 缺损：网络随其缺损程度功能递减。

4. 再学习：比最初学习要快很多（如图 10.5）。

284

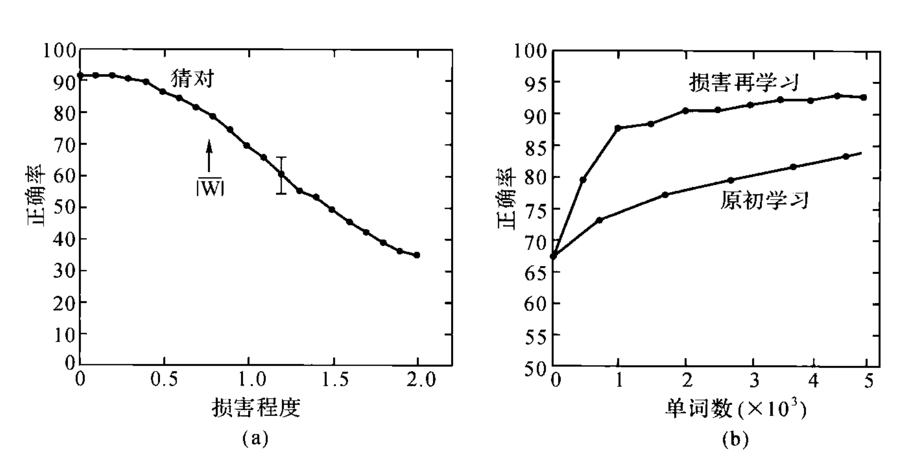


图 10.5 （a）缺省学习曲线和（b）再学习曲线（来源：Sejnowski and Rosenberg，1987：154，图 5）

词典单词运行结果

1. 隐层单元包含 0,15,30,60,120 个单元,网络的运行结果随着单元数量的变化而变化(如图 10.6)。例如,当隐层单元为 0 时,其正确率为 82%;当为 120 时,达到了 98%。

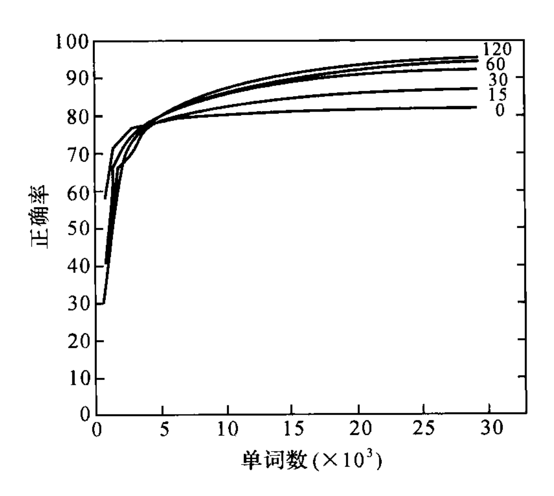


图 10.6　隐层单元数目与学习关系曲线(来源:Sejnowski and Rosenberg, 1987:155,图 6)

285

2. 综合/总的输出正确率:

(i)输入 1000 个单词时,包含 120 个隐层单元网络的正确率为 98%。当有 20012 个单词时:

第一次循环,平均正确率为 77%;

第一次循环结束时,正确率为 85%;

5 次循环后,正确率达到 90%。

(ii)使网络具有双层隐单元层,每层包含 80 个单元:

第一次循环后,平均正确率为 87%;

循环 55 次后,正确率达到 97%。

总的来说,120 个单元的单隐层 NETtalk 与每层包含 80 个单元的双隐层 NETtalk,它们的运行结果相似。

3. 隐层单元分析:在词典任务中,当隐层有 80 个单元,正确率达到 95%时,可用其激活程度验测单词中的不同字母:

(i)对于每个输入,平均会有 20%的单元(16 个)激活,所以 NETalk 不完全是“定位式”,也不完全是“全息”的系统。

(ii)从层级聚类分析(Hierarchical cluster analysis,HCA:相似的项目构成项目组,相似的项目组构成项目群,相似的项目群递级构成更高阶的项目群)中,可以看到元音和辅音的完全分离,以及它们的递阶子划分(如图 10.7)。将与 HCA 相同的程序用于三个工作网络,这三个工作网络起始于三种不同的随机状态。运行结果表明,即使它们的权值模式完全不同,也会得到相同的聚类层次,作者(Sejnowski and Rosenbery,1987:159)因此认为:"单元功能聚类"是这种网络的固有属性。 286

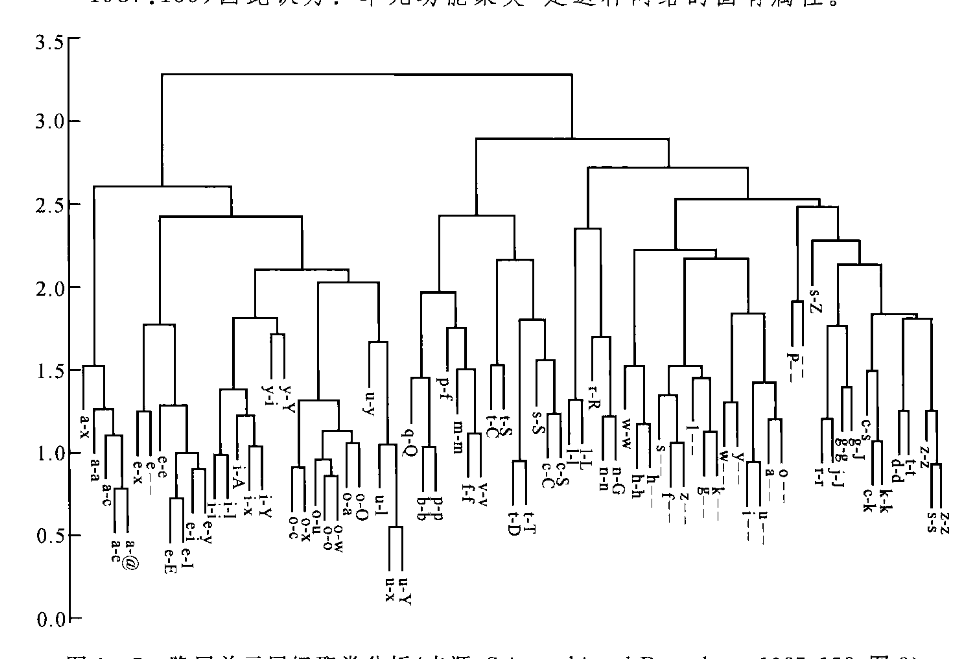


图 10.7 隐层单元层级聚类分析(来源:Sejnowski and Rosenberg,1987:158,图 8)

似乎表明,隐层单元能够学习识别元音和辅音(以及类似的更精细的划分类别)。但有人(Clark,1993,第 4 章)认为,网络并不能用 HCA 表征例如元音与辅音的读音类别,并不能因此认为它掌握了有关元音和辅音的**概念**。隐层单元的优点是使机器的表征指令不受限于程序员本人的有关刺激分类的想法,能更自由地提取出它所能找到的任何规律性。这就意味着一些隐层单元,用"人的"分类图式能给予清楚的解释,而另一些则不能。

注释

[1] 这仅仅是网络总共 68 个单元中的一部分,且与原著并非完全一致。 287

[2] 虽然在数字计算机上实现的网络模拟是离散的,但因片段足够小,仍可接近于模型想要达到的连续性。

[3] 这张图错误地显示了 Lance 是单身,所以 Lance 是"典型的"Jet 成员。但在完整的网络中,Lance 为已婚。

【思考题】

Jets & Sharks(JS)

JS 的结构称作什么?

JS 的表征图式的种类称作什么? 为什么?

JS 具有的三个运行特征是什么?

JS 激活传递原则是什么(使用日常语言表达)?

JS 中的箭头表征什么?

怎样使单元间相互联系,使它们包含在相同的组内或者"组群"中?

JS 如何输入?

JS 如何输出?

JS 怎样从输入计算输出?

NETtalk(NT)

NT 的结构称作什么?

输入节点表示什么?

输入节点的连接表示什么?

输出节点表示什么?

输出节点的连接表示什么?

出于哪种考虑,可以说 NT 的输入和输出层同时是"定位"和"分布"的?

NT 的隐单元层对于表征有什么新的发现——是怎样编码的?

288 NT 怎样进行计算?

NT 如何编程(忽略阈值)?

简述有关 NT 学习的基本观点。

NT 的哪些运行特征与人类行为相似?

【推荐读物】

在 McClelland and Rumelhart (1988)的影响下,"Jets and Sharks"已经被尝试用于家庭计算机中。Bechtel and Abrahamsen (1991)的第 2 章,以及 Clark (1989)的第 5.3 节,包含了有关这一问题的进一步细节。有关 NETtalk 的讨论可在很多认知科学教材,以及 Clark(1993)中找到。在 Verschure (1992)中还有更进一步的阐释。

11 联结主义:基本概念与种类

11.1 引　言

在这一章中,将更加系统地阐述联结主义机器的一些基本概念。我们 289 将检验一些运用这些概念构建的各种不同类型的机器,以及它们的共同特征。能够发现对这种当代快速发展的模型的思考,可以一直追溯到许多世纪以前。在前面,已经介绍了经典联想主义(第 1 章),巴甫洛夫与条件反射(第 2 章),拉什利和赫伯(第 3 章),感知器(第 4 章),以及相互间的争论(第 6 章)。从中甚至可以看到联结主义的库恩范式转换,即从数字、串行、符号到模拟、并行和亚符号的转换(参见 Schneider,1987)。

11.2 基本概念和术语

理想情况是,所有的这些概念应该同时提出,因为它们最终需要依赖彼此间的相互阐释,才能获得完整的理解。然而这是不可能的,所以我们将以一种特殊的顺序提出这些主题。但需要注意的是,只有在一种完整的神经网络框架提出和得到检验之后,才能明白所有这些概念是如何整合在一起的。

连接和权值

我们首先从各个单元之间的连接开始介绍,这些连接被认为是与理想化了的轴突、树突和突触间的连接所作的一种类比,用线段或者“棱脊”表

示。每个连接均有明确的方向，负责传递兴奋（箭头或正数表示）或抑制（实
290 心点或负数表示）。最后，最重要的是每个连接均有一个“权值”或“强度”，由一个具体数值表示，表明这个连接传递了多少激活作用（如果是正数表示兴奋，而如果是负数则表示抑制）。一般而言，权值的记法表示的是从一个单元到另一个单元的连接强度。但与通常直观的理解不同，如：W_{21}，表示的是从单元1到单元2的连接权值，而不是反过来。图11.1是简化的单元连接，在这个例子中，单元1向单元2传递了半个(0.5)激活作用，权值为正表示**兴奋**，为负表示**抑制**。

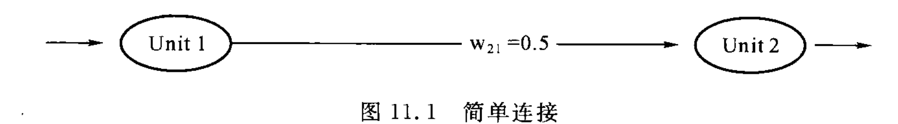


图 11.1　简单连接

单元和激活

单元，或者从环境（“输入单元”）中获得激活作用并传递给其他单元，或者从其他单元获得激活作用，再传递给环境（“输出单元”），还可能从其他单元获得激活作用再传递给另一些单元（“隐层单元”）。“单元”或“节点”可以看作是理想化的神经元胞体。单元用圆圈表示，由一条线段引入（输入），或者一条线段引出（输出），或者两者兼有。

每个单元都处在某种激活状态：正值、零或负值。每个单元对所有向它输入的单元的激活值进行加和（“群收”），每个单元同样可以将激活值传递给所有与它有输出关系的单元（“群发”）。加和的输入值称作“**净**”输入激活。计算净输入值，最简便的方法是把各个输入单元的独立输入值相加。这个净激活值再与先前单元的激活值相加，形成单元（新的）**当前激活**状态。为求简化，在一些例子里我们使新的激活状态等于净激活值。之后这个单元把激活值传递给其他单元，就是**输出**激活。因而在一个单元中存在三种过程：净激活、当前激活以及输出激活（见图11.2和图11.3）

291

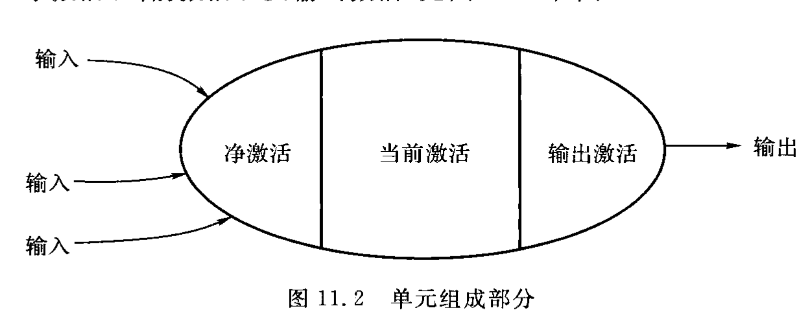


图 11.2　单元组成部分

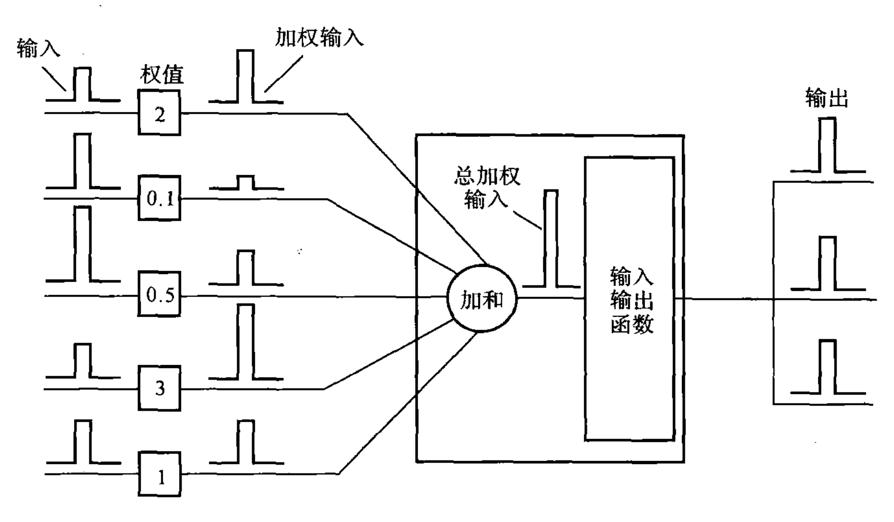


图 11.3 单元连接(来源:Hinton,1992:146,图 10.2;经许可复制)

激活作用从一个单元到另一个单元的传递形式可以视作一种对它们的约束作用。当整个联结网络运作时,每个单元都试着满足其单向输入单元的竞争约束。当整个联结网络满足约束条件的数目不再增加时,就进入了稳定状态。

计算净激活值、当前激活值以及输出激活值有很多规则,我们稍后提及。现在我们介绍一种最简单的规则,即每个单元的当前激活值与输出激活值等于净输入激活值,单元的净激活值等于所有群收之和,每个输入值为上一个单元的输出值与它权值的乘积:

(N)

净输入激活值=群收之和

292

这样,整个过程就包括接收与其连接单元的加权输入,将所有加权输入加和形成净输入,以及将净激活输出。

编程和计算

对网络进行编程,就是设定它的**激活规则**,**阈值**(如果有)以及对每个连接赋予**权值**,可由程序员完成,或者网络通过训练自身获得某组权值。当机器以第二种方式配置的时候,我们就说它已经学会了如何获得权值,而不是通过程序员编程完成。但一旦通过学习获得了某权值组,之后它还会将这组权值应用于另一个网络。稍后我们再回到这个问题。

示例

单元C从两个单元获得输入：单元A通过连接传递了0.7的激活值而增强了0.3倍；单元B通过连接传递了1.0的激活值而增强了0.5倍(图11.4)。对于C来说，净输入值是多少呢？使用公式(N)，首先必须对每个单元(A,B)的输出乘以它们与单元C的连接权值。对于A来说是0.7×0.3=0.21，对于B为1.0×0.5=0.5。之后将这两个数值相加得到：0.21+0.5=0.71。这就是C的净激活值。

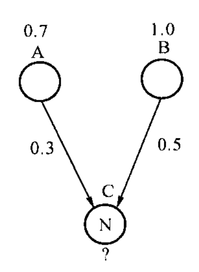


图11.4 网络示例

简单网络

293 思考一个简网络的例子，如图11.5。这个网络有四个单元(A,B,C,D)；每个输入单元将激活作用通过连接传递到输出单元。但为了能够计算，我们需要对它设置传递规则和权值(?)，给它某种输入(??)，计算其输出(???)。现在通过(1)为每个连接设定权值，(2)指定每个单元在下一步中接收和传递激活作用的计算规则。

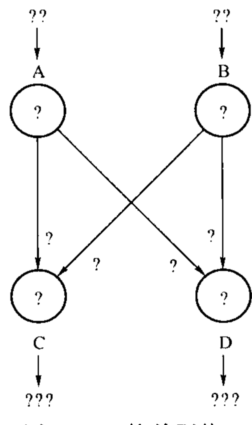


图11.5 简单网络1

联结权值

A－C＝0.1;A－D＝0.3;B－C＝0.2;B－D＝0.4

激活

令净激活值为加权输入之和——公式(N),令当前激活值等于净激活值,输出值等于当前激活值。在这个简化的系统中,净激活值通过(N)在每个单元中传递(但并不总是这样,回忆前面有关 NETtalk 网络的讨论),设定后的网络如图 11.6。

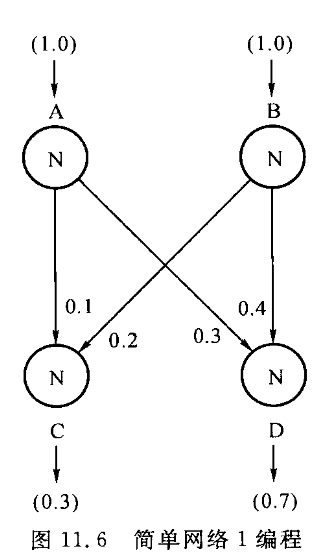


图 11.6　简单网络 1 编程

计算

现在给定这个网络输入,然后计算其输出。输入:让两个输入单元都获得 1.0 的激活作用——图上部括号内说明。单元 A 获得 1.0 的激活作用,因而它的净激活值就是 1.0(参见公式(N)),因此它传递的激活值也是 1.0。单元 B 也是如此。单元 C 同时从 A 和 B 得到激活作用,因此根据公式(N),C 的净激活值这样计算:从每个单元输出的激活值乘以其连接的权值。对于 A 来说为 1.0×0.1＝0.1,B 是 1.0×0.2＝0.2。根据公式(N),将这两个数值相加得到 C 的结果 0.1＋0.2＝0.3,即: 294

A 净输出: 1×0.1 ＝0.1
B 净输出: 1×0.2 ＝0.2
　　　　　　　和 ＝0.3

因此单元 C 的净激活值是 0.3。由于输出值等于净激活值,输出值也是 0.3。单元 D 也从 A、B 获得了激活作用,因此它的计算也遵照同样的方法:

A 净输出: 1×0.3 =0.3
B 净输出: 1×0.4 =0.4
和 =0.7

295 因此,D 的净激活值是 0.7。由于输出值等于净激活值,输出值也是 0.7。因此在这个简单网络 1 中,依照程序设计,C 和 D 括号里的为输出值,A 和 B 括号里的为输入值。

模式联结器

我们已经解释了在简单网络中的计算。下面这个有时称作“模式联结器(Pattern Associator)”的网络,也用来阐释网络的计算。但我们也将用它阐明联结网络的几个附属特征:兴奋与抑制、模式联结、缺省输入以及叠加输入模式。

结构

思考图 11.7 中的简化网络,它共有 8 个单元(4 个输入和 4 个输出),分列成两层。上一层的每个单元都与下一层的每个单元相连接(Rumelhart and McClelland,1986a:33-40)。

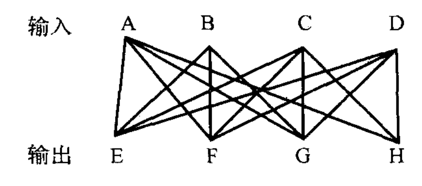


图 11.7 简单连接网络

编程

与前面相同,令每个单元的输出值等于其净输入值(N)。每个连接设定某一权值,一些是兴奋(+),另一些为抑制(-)。当网络非常复杂的时候,这些设置很难像前面一样简明地展示出来,因此需要引入一种更加清晰的矩阵模型。这两种编程技术是等价的,可以按照规则从一个等价转换到另一个(如图 11.8)。

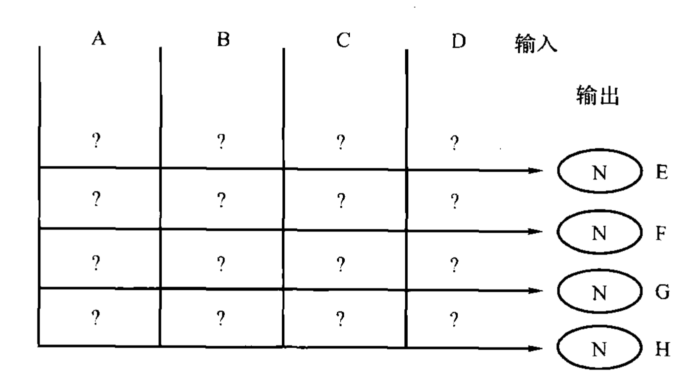


图 11.8　简单网络连接矩阵

问号表示权值,为问号设定不同的数值,会得到一种完全不同的网络。我们以图 11.9 所示矩阵为例,通过设定数值,使网络可对玫瑰外观(输入单元的激活)与玫瑰香味(输出单元的激活)间建立起联系。 296

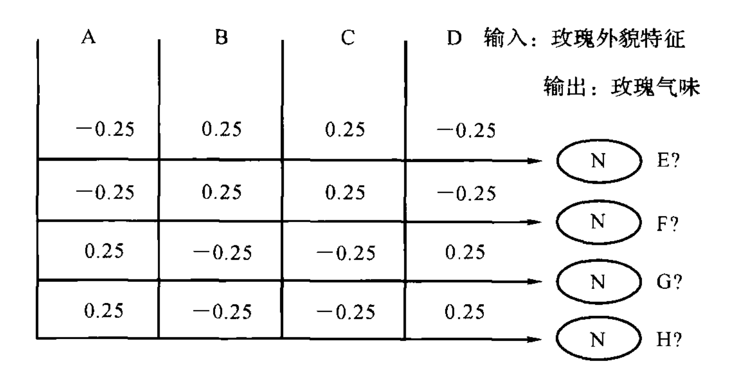


图 11.9　玫瑰网络

假设玫瑰外观可以分解为四个基本部分:A,花瓣的形状;B,花瓣的整体构造;C,花瓣的颜色;D,花茎。我们同样假设它们与玫瑰香味之间存在真实的对应关系(虽然尚为得到科学或香水产业的证明)。按照这种设计,令每一个节点的特定激活值代表玫瑰的某一**具体**的组成部分。例如,节点 A 的激活为"1",代表花瓣的形状;节点 B 是"－1",代表花瓣的构造;节点 C 为"－1",代表花的颜色。其他的输入和输出节点也是如此。

计算

设定每个输入单元的激活值后,就可以计算出每个输出单元的激活值。 297
例如,设定输入节点(A,B,C,D)的值称为("输入矢量")为〈1,－1,－1,1〉:

A:1

B:－1

C:－1

D:1

根据前面给出的公式(N),每个输出单元的激活值就是四个加权输入值之和,如,E 的输出:

连接	权值
A－E:	1×－0.25 ＝－0.25
B－E:	－1×0.25 ＝－0.25
C－E:	－1×0.25 ＝－0.25
D－E:	1×－0.25 ＝－0.25
	和 ＝－1.00

因此单元 E 的净激活值是－1,由于它的输出等于净激活值,所以它的输出也是－1(见玫瑰网络中的括号内说明)。现在我们可以说玫瑰网络产生了单元 E 的输出结果,其输入矢量为〈1,－1,－1,1〉,输出值为－1。F,G,H 的输出也遵循相同的计算,得出下面的值:

E:－1

F:－1

G:1

H:1

因此,玫瑰网络的输出矢量为:〈－1,－1,1,1〉。

练习 1 根据公式(N),运用计算 E 值的方法证明 F,G,H 的值。

玫瑰网络按输入矢量〈1,－1,－1,1〉计算后得出输出矢量〈－1,－1,1,1〉,就是将输入矢量转换为输出矢量。这种转化就是将输入矢量乘以连接权值(在玫瑰网络中用权值矩阵说明),然后将之加和。这就是联结网络的基本计算方法,换个(传统的)说法就是,使输出矢量与输入矢量产生**联结**。如果输入矢量表示比如说是玫瑰的直观特征,输出矢量表示玫瑰的香味,那么玫瑰网络就实现了玫瑰的外观与其香味之间的联系——可以通过玫瑰的外观知道玫瑰的香味。

298

缺省输入

有趣的是,如果给网络一个错误或缺省的输入(不输入),它仍能够进行

有效计算——网络并不会因此而瘫痪。举个例子,假如玫瑰网络中单元 C 并没有输入激活,它的输入也就是‘0’(因为 C 是颜色检验器,这个玫瑰网络可能会显示一张玫瑰的黑白照)。在这种情况下,我们就需要以新的输入矢量〈1,－1,0,1〉,重新计算单元的输出。再以单元 E 的输出为例,按照公式(N)计算得到:

连接	权值
A－E:	1×－0.25 ＝－0.25
B－E:	－1×0.25 ＝－0.25
C－E:	0×0.25 ＝0
D－E:	1×－0.25 ＝－0.25
	和 ＝－0.75

因此,经过对新的缺省输入矢量〈1,－1,0,1〉(如缺少颜色)的计算,E 的输出激活值为－0.75。同样地计算 F,G,H 单元的输出,得到:

E:－0.75
F:－0.75
G:0.75
H:0.75

练习 2 证明 F,G,H 的输出结果。

可见,玫瑰网络将缺省输入矢量〈1,－1,0,1〉转化为输出矢量〈－0.75,－0.75,0.75,0.75〉。对比正常数据与缺省数据:

正常:〈1,－1,1,1〉
缺省:〈－0.75,－0.75,0.75,0.75〉

这表明,虽然输出结果不同,但符号模式是一样的,正如鲁梅尔哈特和麦克莱兰德(1986a:35)所说:"……模式作为系统的反应……(输入正确的数据符号)会准确地激活所有单元;即使有时其准确性会出现某种程度的削弱,但其完整的模式……还是会呈现出来。" 299

叠加网络

联结网络还有一个重要的,也有一点奇怪的属性,就是它们可同时储存不止一对输入－输出矢量间联结的能力。为说明这一点,我们构造另一个联结网络,沿着同样的线路,如玫瑰网络一样,但联结的是山羊状貌与山羊

气味(如图 11.10,也可见 Rumelhart and McClelland,1986a:35)。同样的,我们假设输入节点得到不同的激活数值可表示山羊的状貌,输出节点的激活值表示山羊的气味。正如括号里说明的数据,这个网络将山羊状貌——输入矢量〈—1,1,—1,1〉,转换为山羊气味——输出矢量为〈—1,1,1,—1〉。

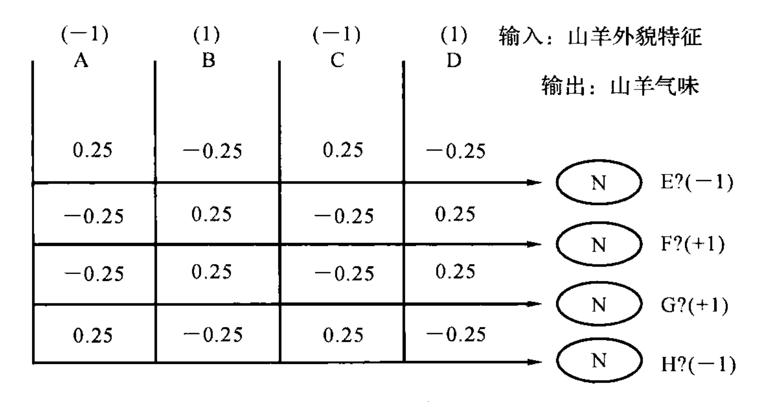


图 11.10　山羊网络

练习 3　证明山羊网络的输入矢量〈—1,1,—1,1〉与输出矢量〈—1,1,1,—1〉的联结。提示:将 0.25 累加到 1,即可证明。

我们是否能构造一种单个的"玫瑰—山羊"网络呢,它只依据玫瑰的外
300
观显示玫瑰的香味,只根据山羊的状貌显示山羊的气味?(倘若我们已经知道答案不是"能够",似乎没必要做这项工作了。)首先需要阐明这样一个基本事实,即两个权值矩阵,如嵌入玫瑰网络及山羊网络中的两个权值矩阵,可以相加。它们相加在一起,最终的矩阵仍会产生正确的输出(不会有山羊气味的玫瑰)。那么怎样将矩阵相加呢?答案是需要将它们相对应部分相加:

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix} + \begin{bmatrix} w\ x & \\ y & z \end{bmatrix} = \begin{bmatrix} a+w & b+x \\ c+y & d+z \end{bmatrix}$$

在玫瑰—山羊网络中我们将原来两个权值矩阵相加,结果矩阵网络如图 11.11 所示(参见 Bechtel and Abrahamsen,1991:49-50)。

计算

现在需要证明,玫瑰和山羊的联结网络不会使山羊的气味像玫瑰(反之也不会)。设定在这个网络中,山羊状貌矢量为〈—1,1,—1,1〉,看看会得到

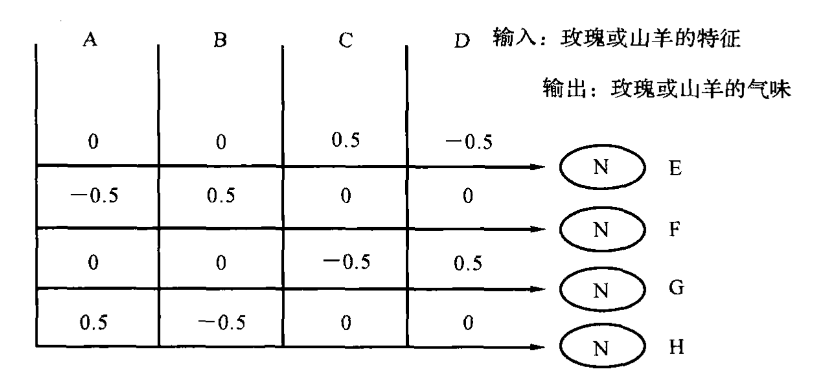


图 11.11　山羊一玫瑰网络

什么样的结果。同前边的计算一样，首先详细地计算第一个输出单元 E 的结果：

连接		权值
A—E：	−1×0	=0
B—E：	1×0	=0
C—E：	−1×0.5	=−0.5
D—E：	1×−0.5	=−0.5
	和	=−1

到目前为止一切顺利，但山羊气味的矢量和玫瑰香味的矢量都是以−1 开始(输出单元 E 中，两个矢量都是以−1 为起始输出值)。因而，为了确保我们是在正确的轨道上，应该计算一个山羊和玫瑰气味值不同的输出单元。正如我们看到的，对于玫瑰 F=−1，但是对于山羊 F=1。那么对于 F 单元，从玫瑰一山羊网络中会得到怎样的山羊矢量呢？　301

连接		权值
A—F：	−1×−0.5	=0.5
B—F：	1×0.5	=0.5
C—F：	−1×0	=0
D—F：	1×0	=0
	和	=1

单元 F 的输出值 1，对应山羊状貌的输入矢量，这是正确的。

练习 4 证明玫瑰—山羊网络中，山羊气味对应山羊状貌，玫瑰香味对应玫瑰外观。

技术性补充 输入矢量具有某些限制，网络才能够实现同时储存。这些限制部分地与网络包含单元的个数，对于网路而言输入矢量间是否分离，以及应用的算法类型相关。两个输入矢量间必须具备哪种关系，才能够使联结网络能够同时学习这些矢量，而不相互产生干扰？两个矢量必须为正交矢量。当两个矢量 V1，V2 彼此处于垂直关系时，那么它们就是正交的。也就是说，它们的内乘积(数量积)＝0。

例如，令 V1＝⟨－1，1，1，－1⟩，V2＝⟨1，1，1，1⟩。V1 和 V2 之和＝0，即－1×1(＝－1)＋1×1(＝1)＋1×1(＝1)＋－1×1(＝－1)＝0，它们的关系就是正交。在一个 n 维(单元层)系统中，至少存在 n 个相互正交的矢量。

递归网络

在这节中，我们前面介绍的都属于顺向网络，也就是说激活作用在网络中从输入到输出没有回溯环路。如果引入这样的环路，网络可以循环激活(可以将之想象为形成简单记忆——回顾第 3 章赫伯回响环路)，这样网络便随时间变化而产生了某些演化现象。在一个网络中，即使只局限在三层网络，也有很多种反馈激活的方式。例如，指定单元层上的不同单元可以反馈到其他不同单元层上，或者也可以要求在同一层上所有单元向它自身所在单元层反馈。可以说“初级的(三层)递归网络”，要么(1)高层向低层反馈，要么(2)指定层向自身反馈(见图 11.12)。

302

而“扩展的(三层)递归网络”则包含一个独立的单元扩展层，可以向(a)隐单元层，或者(b)输入单元层反馈(有时这种区分的术语在各种文献中常常被忽略)。艾尔曼(Elman，1990a，1990b，1992)提出了一种著名的扩展递归网络。在他的网络中具有他称作“背景层”的单元层，能够从隐单元层得到既定(未修正的)连接，然后再以修正的连接反馈给隐单元层(见图 11.13)。

这个扩展“背景层”能够储存网络前一步的激活，也简单地储存了前一步激活的效果。从背景层中获得这样的信息，隐单元层就有了先前活动的记录；在 t＋1 时刻，输入隐单元层的信息中同时包含 t 时刻的信息；t＋2 时刻的输入信息包含了 t＋2，t＋1 和 t 时刻所有信息。递归网络可以运用这

些信息执行随时间扩延的任务——如果在网络中能够找到连续且独立的数据。

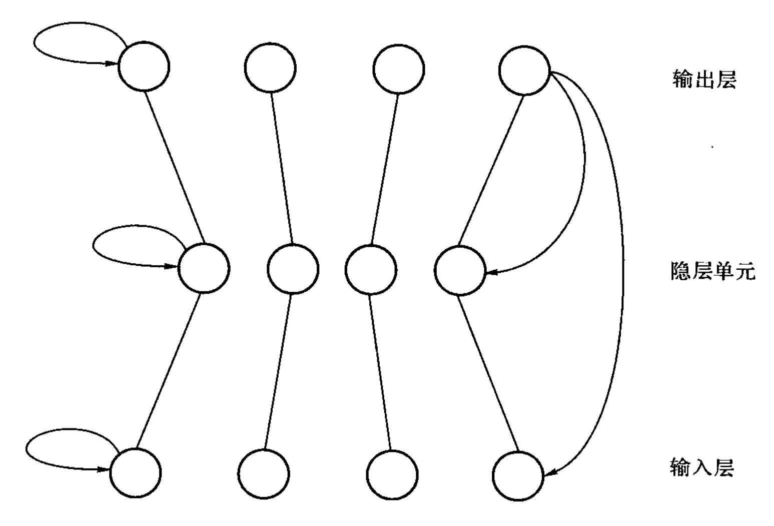


图 11.12 非扩展递归网络图式

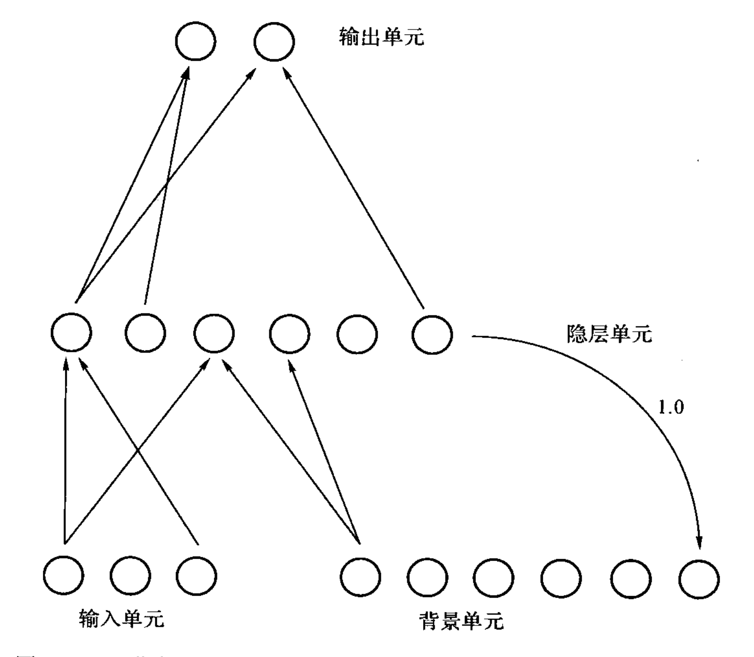


图 11.13 艾尔曼(扩展)递归网络结构(来源:Elman,1990b:349,图 1;经 MIT Press 许可复制)

递归网络以前向三层网络（通常是反向传递，如 NETtalk 所运用的）相同的算法训练，艾尔曼（1990a）说明了递归网络如何发现各种不同领域中的序列规则。

字母预测

例如，艾尔曼构建的一种递归网络，包括 6 个输入单元，20 个隐层单元和 6 个输出单元，以及 6 个背景单元。基于以下规则构建了字母表：

> 1. 首先，三个辅音字母 b，d，g 随机组合，获得一个包含 1000 个字母的序列。
>
> 2. 然后，用下面规则，替换辅音。在 b 后面引入元音 a，d 后面引入元音 ii 以及 g 后面引入元音 uuu：
>
> b→ba
>
> d→dii
>
> g→guuu

304 这样就产生了一个辅音和元音串，如 diibaguuubadiidi-iguu... 既然起始辅音的顺序是随机的，那么任何学习机器都不能预测它们的具体位置，因而对每个辅音预测的错误率很高。但由规则引入元音序列 a，ii，uuu 后，就可以完全预测了。就是说，一旦任何发现序列中的 b，d 或者 g，机器能够知道它后面紧接着的是哪个元音以及元音的个数。递归网络就是这样进行字母预测的。在训练 200 次后，基于原始序列的遵循三个元音规则的新序列会被机器提出。但也如预期一样，辅音产生了随机的错误，但是机器仍能预测元音以及它们的个数。

字母一单词预测

艾尔曼调整了辅音和元音串的顺序，形成 15 个单词（单词之间不存在间隔）。有 200 个句子，每个句子中包含这些单词中的 4 到 6 个。这样 200 个句子就排列成了包含 1270 个单词的序列。单词分解为字母，就是一个包含 4963 个字母的字母串。设置一种包含 5 个输入单元，20 个隐层单元和背景单元，以及 5 个输出单元的递归网络。把字母串向这个网络完整地呈现 10 次，这个次数对于网络来说太少了，因此它不可能记住整个序列。测试时，尽管产生了一些错误，但网络已经能够识别重复出现的“类单词”字母组。开始时，网络对某一单词的起始字母有很高的错误率（它并不知道后面接着的单词是什么），但一旦它识别了这个单词的起始字母，预测以那个字母开

头的单词,就会变得非常容易了(见图 11.14)。后来艾尔曼还将网络扩展到识别口语中简单句子中的单词,之后又识别了复杂句中插入从句的单词。

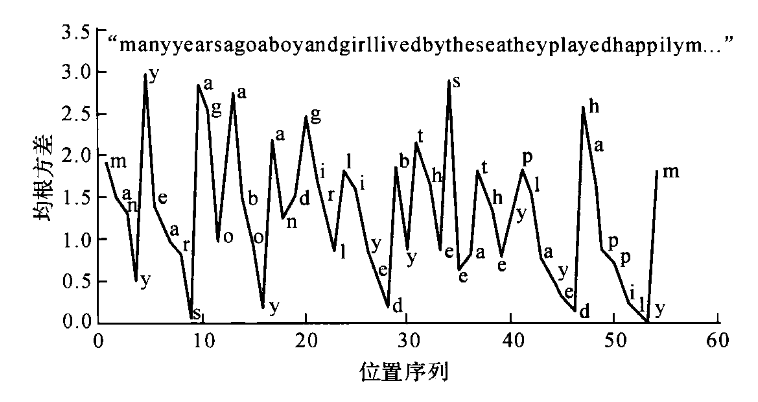


图 11.14　预测字母—单词任务误差(来源:Elman,1990a:194,图 6;经作者许可复制)

11.3　学习与训练

一般而言,在联结网络中的学习(或“训练”)是有着清楚的应用限制的。网络的整体结构并不会受影响(特例见下),激活传递规则也不受影响,对单元和矢量的解释也如此。基本上,“经验”产生的影响是使连接权值发生变化。虽然如此,还是需要注意某些权值的改变仍可看作其结构发生了改变。305
例如,如果某一权值设定为 0,它的功能就等同于没有连接。

(1)根据训练所使用的范型(paradigm)分类学习程序(参见 Rumelhart and Zipser,1986:161):

> *规则检测器*(Regularity detector):在这种范型中,存在着一种刺激样式种群,每种刺激样式均有出现的几率。系统需要发现那种种群刺激样式在统计学上的显著特征。因为并没有预先设定样式的分类方法,因此系统需要形成它自己的对于输入刺激的特征表征,采集到种群刺激最显著的特征。
>
> *自动联结器*(Auto associator):在这种范型中,一组样式重复出现,系统需要将这些样式储存。之后,如果原始样式中的某部分或一个与原始样式相似的样式出现,系统可通过完备样式程序重新获得原始样式。

306

样式联结器(Pattern associator):自动联结器范型的一种变型。样式组对重复出现。系统能够学习当组对中的一个出现,也会将组对中的另一个呈现出来。

分类范型(Classification paradigm):分类范型可看作是基于前面的范型而提出的一种新的范型。在这种范型中,系统能够按照一套固定的分类标准对刺激样式予以分类。在训练环节中,刺激样式以及刺激所属类别呈现给系统,目标是让系统学会正确的刺激分类方式。从而如果以后出现了一种特殊刺激或者有些细微歪曲的刺激,系统也能够正确地给予分类。感知器就是这种范型的早期例子。

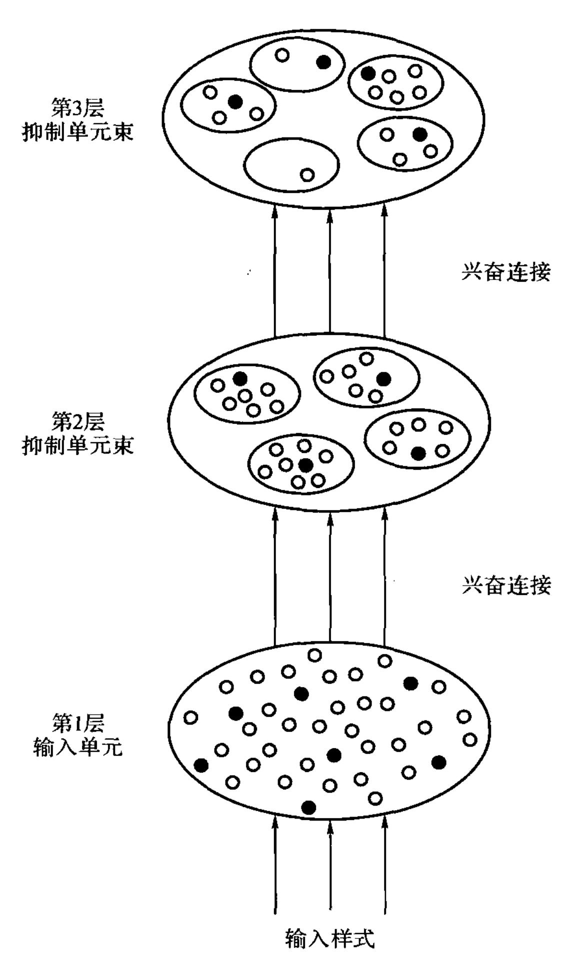
(2)另一个学习程序的分类方法,是根据它们是否使用了一个“老师”——如果是,称为“监督学习”;如果没有,则称为“无监督学习”。处于两者之间的是“强制学习”,系统直接操作输入刺激的内容,而不是根据输出调节权值。我们主要关注监督学习,但首先大致了解一下无监督学习。

无监督学习

在无监督学习中,网络并没有通过训练而要近似接近的输出目标。一种已受到充分研究的形式是“竞争学习”。竞争学习在种群刺激中,通过对显著样式形成内部表征检测刺激规则(见图 11.15)。竞争学习网络将单元层按等级排列为不同层次,层次之间存在兴奋作用,而抑制作用则发生在每层单元束内部。随着系统的学习,一些单元开始对相同的刺激产生反应,而不同的刺激也有不同的单元与之对应。

监督学习

监督学习需要给网络某一序列学习对,一个学习对包括一个输入矢量和一个目标矢量。输入矢量是应用矢量,网络基于输入矢量计算输出矢量。训练算法将实际输出与目标输出进行比较,然后根据一些具体规则调整相关的连接权值。不同的学习程序使用不同的规则。



竞争学习机制的结构。竞争学习发生在按等级排布的单元集中,单元在这个图解中用圆点表示。单元可以被激活也可以不被激活,激活的单元用实心圆点标识,不激活的标识为空心圆点。一般而言,指定层级中的单元,从低层级的所有单元中获得输入,对其高层级的所有单元投射输出。层级间的连接为兴奋,层级内部单元间的连接为抑制。每一层级中包含着几组互相抑制的单元束。一个单元束内只有一个单元为兴奋,以这样的方式实现单元束内单元的彼此抑制。我们认为特定单元层内,单元激活的构造为其下一高阶单元层的输入样式。有多少个这样的单元层是可随意安排的。虽然特定单元束内包含固定数量的单元,但不同的束内可以有不同数量的单元。

图 11.15 竞争学习机制结构(来源:Rumelhart and McClelland,1986:163,图 2;经 MIT Press 许可复制)

赫伯学习

308　我们曾在第 3 章提到，赫伯(1949)提出了一个非常重要的观点，认为当两个相连接的神经元同时兴奋时，它们之间的连接强度会增加[1]。这就产生了两个问题：其一，将这种观点应用于学习或训练的网络，它属于监督学习还是无监督学习呢？其二，连接强度会增加多少，受什么决定？对于第一个问题，当涉及任意两个单元时，就像赫伯说的，连接"权值"会有所调整但没有"监督"。但以简单联结器为例，第二个(B)"细胞"可能就是输出节点，因为权值的变化依赖它的输出，所以调整与它的连接就可有效地进行"监督"。因此对于这种网络，说它不属于"监督"网络，是值得商榷的。我们把赫伯学习看作是一种特殊的监督学习。对于第二个问题，人们已经提出了很多观点，可使这个问题变得更为精确。我们将采纳如下的观点(参见 Bechtel and Abrahamsen，1991：48ff，72ff)：权值的变化量等于第一个单元的激活值乘以第二个单元的激活值，再乘以学习速率(系统权值改变的速度)。在简单网络中，如我们前面介绍的，第一个单元是输入单元，第二个单元是输出单元。通常学习速率是个分数，输入单元的个数为分母，1 为分子。因此可将赫伯学习规则概括如下：

(H)

(1)计算输入单元激活值和与其连接的输出单元的目标激活值，以及学习速率(1/输入单元个数)的乘积。

(2)将(1)的计算结果与先前的连接权值相加。

(3)所得结果即为新的连接权值。

以某一简单网络(见图 11.16)为例。如有两组输入矢量，为能够使它们有所区分，设其中一组的输出结果为 1，另一组为 0。为了加深理解，我们假设第一组输出为女性面孔代码，第二组是男性面孔代码。

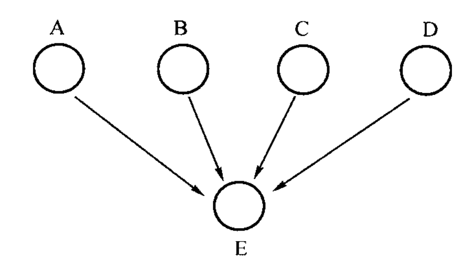


图 11.16　面孔识别网络

女性组(输出=1)

矢量 1:〈1,-1,-1,1〉

矢量 2:〈1,1,1,1〉

男性组(输出=0)

矢量 3:〈1,-1,1,-1〉

矢量 4:〈-1,-1,1,1〉

我们能不能训练这个"面孔网络"使它能够识别这两组脸型呢?训练网络就是要调整它的连接强度,使之能够当输入女性面孔矢量,网络输出为1,输入男性面孔矢量,网络输出为0。每个矢量都有4个连接权值要考虑,根据(H),必须通过下面两个步骤调整每个连接:(a)首先计算输入激活值(输入矢量)与期望输出激活值(女性为1,男性为0)以及学习速率(1/输入单元的个数=1/4=0.25)的乘积;(b)然后将这个值与先前连接权值相加——就是新的权值。 309

矢量 1

我们首先以矢量1,即第一个女性矢量,训练第一个连接A—E。目标是希望网络能够学习当输入矢量1:〈1,-1,-1,1〉时,其输出为1(见图11.17)。

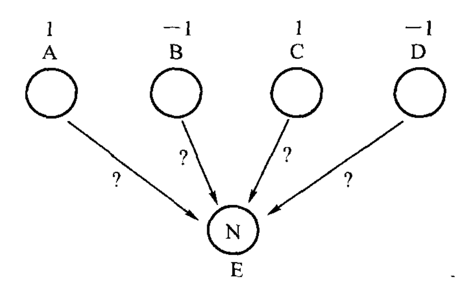


图 11.17 训练前的面孔识别网络

(应用H)

(i)计算输入单元A的激活值,和与其相连接的输出单元E的激活值,以及学习速率(1/4)的乘积。输入单元A从矢量1中获得1的激活值,与其相连接的输出单元E目标获得激活值1,学习速率是0.25,因而结果为1×1×0.25=0.25。

(ii)将0.25与先前的连接权值0相加。

(iii)得到新的连接权值:0+0.25=0.25。

310 因此，对面孔网络矢量 1 的训练结果，是指派了第一个连接的权值 0.25。对其他连接重复相同的计算得到：

连接	旧权值	新权值
A—E：	1×1×0.25+0=	0.25
B—E：	−1×1×0.25+0=	−0.25
C—E：	−1×1×0.25+0=	−0.25
D—E：	1×1×0.25+0=	0.25

这样面孔网络就对矢量 1 进行了训练（见图 11.18）。

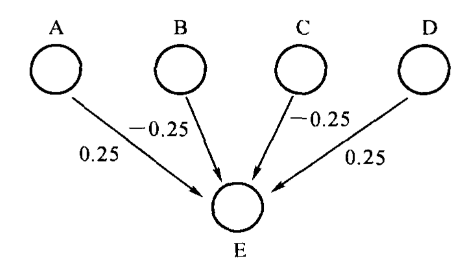


图 11.18　矢量 1 训练后的面孔识别网络

矢量 2

用同样的方法以矢量 2 训练面孔网络。这里唯一不同是步骤（b），此时的先前连接权值是从学习矢量 1 获得的。我们知道，训练目标是使网络学习 311 输入矢量 2：⟨1，1，1，1⟩时，输出为 1，学习速率相同（0.25）。所以我们自然会想到，矢量 2 训练结果的不同在于加或减 0.25 的不同。正是如此，首先计算第一个权值进行验证。

（应用 H）

（i）计算输入单元 A 激活值和与其连接的输出单元 E 的激活值，以及学习速率（1/4）的乘积。输入单元 A 从矢量 2 中获得 1 的激活值，与之连接的输出单元 E 目标获得激活值 1，学习速率是 0.25。因而结果是为 1×1×0.25=0.25。

（ii）将（i）的计算结果与先前的连接权值 0.25 相加。

（iii）得到新的联结权值：0.25+0.25=0.5。

所以面孔网络矢量 2 的训练结果分配给第一个连接的权值为 0.5。对余下的连接重复相同的计算得到：

连接	旧权值	新权值
A—E:	1×1×0.25=0.25+0.25	=0.5
B—E:	1×1×0.25=0.25+−0.25	=0
C—E:	1×1×0.25=0.25+−0.25	=0
D—E:	1×1×0.25=0.25+0.25	=0.5

这样面孔网络就对矢量2进行了训练(见图11.19)。

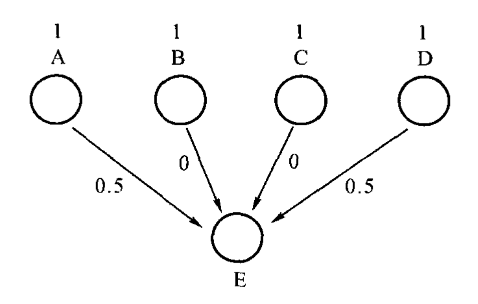


图11.19 矢量2训练后的面孔识别网络

矢量3和矢量4

现在我们对男性面孔矢量3和矢量4训练网络。这就显得非常简单,注意到输出单元的目标激活值在两个矢量中都为0,这就意味着这两个矢量都没有变化。因为运用(H),将目标输出与输入以及学习速率相乘结果为0,任何数字与0相乘的结果都为0。因此增量为0,先前的权值维持不变——这就是面孔网络训练第二组矢量的最终结果了。第一个训练循环至此结束。 312

计算

现在我们来看看所受训练的网络能够做什么——它能够识别男性和女性矢量吗?

矢量2

首先看一下当输入第二个女性矢量时它的输出。矢量2:〈1,1,1,1〉,运用(H)我们得到群收之和:1×0.5+1×0+1×0+1×0.5=1,正确。在一个训练循环后,面孔网络已经能够正确地将矢量2归类于1——女性。

矢量4

接下来看一下当输入一个男性矢量时它的输出。矢量4:〈−1,−1,1,1〉,运用(H)我们得到群收之和:−1×0.5+−1×0+1×0+1×0.5=0,正确。

在一个训练循环后，面孔网络正确地将矢量 4 归类于 0——男性。

练习 5 证明网络能够正确地归类矢量 1 和矢量 3。

技术性补充 在联结网络中，赫伯规则对于学习矢量有哪些限制？遵循赫伯规则训练的线性联结器，只能学习正交矢量才不会出现错误。正如我们在上一个技术性补充中所讲的，只有两个矢量的内乘积为 0 时才为正交矢量。也就是说，将正交矢量相乘，然后加和乘积，会得到结果 0。如果试着教线性联结器一些新的非正交矢量，它的性能将会降低——出现种种错误。

Delta 学习

在赫伯学习中，权值可看作输入激活值与目标输出激活值（还包括学习速率，但学习速率对于网络来说是不变的）的函数。但网络的**实际输出**与网络的**目标结果**之间，可能出现的误差并没有得到反馈，没有利用这种误差产生有益的作用。误差即目标输出激活值与实际输出激活值之间的差值：

313

(E)

误差＝目标激活值－实际激活值（输出单元）

Delta 学习即利用误差信息，当输出单元的目标结果是兴奋，但实际得出的却是抑制时，提高连接权值；当输出单元的目标结果是兴奋，但实际得出的却是兴奋时，降低连接权值；当权值正好合适（符合期望结果），不作变化。Delta 学习的非公式规则如下：

(D)

(1)计算误差(E)。

(2)求积：输入值×误差×学习速率。

(3)将(2)与先前权值相加。

(4)结果即为连接新权值。

将这种学习程序应用于面孔网络，使用如下的输入矢量和目标输出值：

输入矢量	输出
V1:〈1,－1,1,－1〉	1
V2:〈1,1,1,1〉	1
V3:〈1,1,1,－1〉	－1

V4:〈1,—1,—1,1〉　　—1

开始时,网络的所有连接权值都为 0。

测试 1(矢量 1)

与前面一样,首先计算当网络学习矢量 V1 与目标输出值 1 的连接时,第一个连接权值 A—E 的变化。对这个连接,应用(D)得出:

(应用 D)

(i)计算误差(E),即输出单元的目标结果与实际结果之差。目标得出 1,但它的实际结果是多少呢?再次使用(N):输出是各个输入值与权值乘积之和,但此时所有权值都是 0,所以积是 0,和也为 0——实际产生结果为 0。因此误差就是 1 与 0 之差。 314

(ii)求输入激活值与误差以及学习速率之积。输入激活值为 1,误差是 1,学习速率是 0.25,所以它们的乘积为 1×1×0.25=0.25。

(iii)与先前权值相加。原来权值为 0,所以结果为 0+0.25=0.25。

(iv)0.25 就是新的连接权值。

这样我们就得出了连接 A—E 的新权值 0.25,对剩下的连接作相同的计算。显然,除了正负值有所变化外(B—E 和 D—E 是"—"),所有的数值都是一样的。所以,矢量 V1 新权值的计算结果为:

A—E:0.25

B—E:—0.25

C—E:0.25

D—E:—0.25

练习 6　证明后三个连接的结果。

测试 2(矢量 2)

运用(D)以同样的方法来计算矢量 2:〈1,1,1,1〉。从第一个连接A—E开始:

(应用 D)

(i)计算误差,即输出单元的目标结果与实际结果之差。目标得出 1,实际结果是什么呢?需要再次运用(N):输出是各个输入值与权值乘积之和:1×0.25+1×—0.25+1×0.25+1×—0.25=0.25+—0.25+0.25+—0.25=0。因而实际输出为 0。所以误差就是 1—0,即 1。

(ii)求输入激活作用值与误差、学习速率的乘积。输入激活作用值是1,误差1,学习速率0.25,所以它们的乘积为1×1×0.25=0.25。

315 (iii)与先前权值相加。先前权值为0.25,因而结果为0.5。

(iv)新连接权值为0.5。

对剩下的连接进行同样的计算,得出:

A—E:0.5

B—E:0

C—E:0.5

D—E:0

练习7 证明后面三个连接新权值的结果。

测试3(矢量3)

对第三个矢量进行同样的计算,得到新权值(注意目标输出值在这里是—1)为:

A—E:0

B—E:—0.5

C—E:0

D—E:0.5

测试4(矢量4)

对矢量4进行同样的计算,得到的循环一次后的连接权值(注意目标输出值是—1):

一次循环后的权值

A—E:—0.5

B—E:0

C—E:0.5

D—E:0

练习8 证明矢量3和矢量4的计算。

我们已经使网络完成了矢量V1—4的训练,也就是一次训练循环(一次训练循环为完成所有输入—输出对的测试)。还可以使网络进行多次训练循环。例如,进行了20次训练循环后所具有的网络权值(如图11.20所示,也可参见Haberland,1994:185)。

316

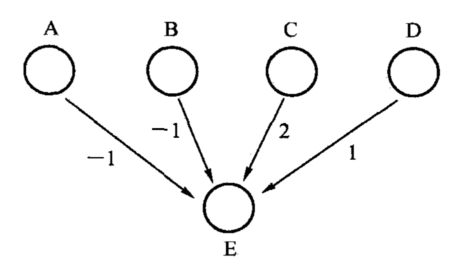


图 11.20　Delta 训练 20 次循环后的网络

计算

可以证明这个网络对这四个矢量都会产生正确的输出。如矢量 3〈1,1,1,－1〉,目标结果是－1,能够得到－1 吗?

(应用 N)

群收之和＝1×－1＋1×－1＋1×2＋－1×1＝－1＋－1＋2＋－1＝－1。

网络对矢量 3 的输入得出与目标值相同的结果。

练习 9　证明网络对矢量 1、矢量 2 和矢量 4 的正确性。

技术性补充　Delta 学习怎样克服赫伯学习的不足呢?如果矢量不是正交的,而是线性分离的,那么赫伯规则将无法学习,但 Delta 规则却可以。即使矢量不是线性分离的,Delta 规则仍起作用:它可将误差降低到最小。Delta 学习的局限是什么呢?Delta 规则只能应用于线性独立的矢量,而不能学习非线性分离的矢量。而加入隐单元层就不需要如此了,这时具有隐单元层的网络应用广义 Delta 规则(具有隐单元层的 Delta 网络是广义的)就非常必要了。

反向传递(广义 delta 规则)

如果存在隐单元层,Delta 规则将无法工作,因为它必须直接连接着输入和输出,这时就需要引用广义 Delta 规则了。广义 Delta 规则也被称作误差反向传递(或者是“反向传递”)。对于输出单元而言,反向传递与 Delta 规则相似——隐层单元与输出层单元之间连接权值的变化与输出单元误差成正比。但是输入单元与隐层单元之间的连接,由于没有目标激活值,因此也就没有 Delta 规则可以使用的误差信号。那么怎么能计算隐层单元的误差从而降低误差呢?这正是反向传递所要做的:它可以将误差总和通过网络　317

反向传递。隐层单元误差与输出单元误差以及隐层单元与输出单元的连接强度成正比。这正是一种确定低层次单元对输出层的误差影响的方法(参考推荐读物)。

技术性补充 在第 4 章中介绍了 XOR 问题是解决非线性分离难题的一个较好示例。正如我们所看到的,XOR 问题在二维空间内不能被一条直线分割开(见图 4.13),这是由于两个完全不同的输入⟨0,0⟩和⟨1,1⟩,需要得出相同的输出,也就是 0。通过引进隐单元层则能解决这个问题。隐单元层能够实现输入结构的转换,使两个完全不同的输入转化成为相同的样式,如⟨0,0⟩。隐单元层网络的这种特征,能够把输入空间转换成线性分离的问题,这在很多认知任务中被证明是非常重要的。

11.4 表 征

绝大多数联结主义者和非联结主义者都会赞同这样的说法,即想要对人类认知能力作充分的解释,就需要将认知看作部分地涉及**表征**"**操作**"(生成、转换和删除)。那么认知功能的联结主义模型就必须能够模拟这个过程。正如我们在第二部分中提到的,应用联结主义模型时存在两个问题:

(Q1)

联结主义模型"操纵"(生成、转换和删除)表征的类别?

(Q2)

联结主义模型的表征如何表征它们所表征的事物——什么决定了所表征的内容?

318 两个问题都被称为表征难题。需要注意的是,表征难题 Q1 中的"表征"为复数(the problem of representations),表征难题 Q2 中的"表征"为单数(the problem of representation)。回答第一个问题需要知道联结主义模型表征的(1)结构和(2)它的主要程式及特征;对第二个问题的回答可以让我们知道(a)在什么的条件下,某事物才可说是一种表征,即表达了某物,以及(b)表征的内容确切地由什么决定。

问题 1:联结表征的种类

在众多的联结主义著作中,对于表征的操作(生成、转换和删除)本质,

提出了很多方案。但由于一些术语比较模糊,而使这个问题变得复杂。为了能够清楚地给予说明,我们采用下面的约定(再次需要注意,既然并不是所有人都赞同这些表达,我们也不必过于严格,尤其是在阐述他人观点时):

几个术语

定位式(定点)表征:单个单元的激活,表示某一范围内的某个元素,如一个具体概念、性质或者个体。

准分布式表征:一组单个单元的激活样式,表示某一范围内的某个元素,并且这些单元并不参与其他表征。

(完全)分布式表征:一组单个单元的激活样式,表示某一范围内的某个元素,并且这些单元也同时参与其他表征。

定位式表征

定位式表征的基本观点是,一个具体单元专用于某一具体“概念”。前面的 NETtalk 网络的输入单元就是这样,一个单个单元专用于表征一个字母,这样的程式是直觉和简单的。定位式表征是非常简明和容易理解的。但作为一种认知模型却有着严重缺陷,并且不能在神经硬件上直接实现。

定位式表征的缺陷

319 因为表征是直接在神经硬件中实现的,因此定位式表征面临的首要难题是**神经死亡**:脑每天都会死亡成百上千的神经元,如果脑采用的是定位式表征,那么每天也必然会消失一些不确定的概念,但是我们几乎从没有注意到一些概念突然消失,比如说“祖母”。费尔德曼(Feldman,1989:75)认为,这个缺陷可将某一表征分布到至少三个到四个神经元中才能消除[2]。第二个缺陷是并**没有足够的神经元**表征人所接触到的所有事物,包括所有的概念以及概念间的关系。例如,视觉系统需要感知六维空间的信息,即使每个维度在最低水平的 10^6 个像点上只处理 10 个值,还至少需要 10^{12} 个神经元——这实在太多了(参见 Feldman,1989)。最后一个缺陷是,每学习一个新的概念,都需要找到一个与先前所有概念关系连接正确的节点,或者重组网络中相关的所有的连接关系。因此,这种程序(经常)在生物学和计算科学中受到质疑(参见 Rumelhart and McClelland,1986a,第 3 章)。

分布式表征

即使我们同意定位式表征满足直觉的特征，然而在神经水平上却不可实现，计算也不充分，因此需要给概念表征添加更多单元解决这个问题。NETtalk 输出单元就是一个显著的例子。在那里，音素表征分布在多个不同特征的单元中，这些具有不同特征的单元能够同时参与表征众多不同音素。

技术性补充　简单程式和它的难题：思考这个简单例子，如何表征一架飞机上的位置点(或者在视线范围内的一个虫子)。有两组可用的单元，一组表示在 X 轴上的位置，另一组表示在 Y 轴上的位置，这样我们或许就能够表征点 2 的位置，如图 11.21。

也就是说，点 2 的位置可以用一对矢量表示：⟨X:0100⟩，⟨Y:1000⟩。如果想要也表征点 3，就必须使矢量变复杂，如图 11.22。

我们可得到：2＝⟨x:0001⟩，⟨y:0010⟩；3＝⟨x:0001⟩，⟨y:0010⟩。

绑定难题(The binding problem)

我们如何区别(图 11.22 中)点 2 和点 3 与图 11.23 中点 2′和点 3′的表征？

320

2′＝⟨x:0001⟩，⟨y:0010⟩与 3 相同(或者看看对角线上的值)。这样的"串扰(crosstalk)"的产生称作绑定难题：怎样将位置都结合起来就像在图 11.22 中一样，而避免图 11.23 中的重像(2′，3′)。

0				
0				
0				
1		2		
y/x	0	1	0	0

图 11.21　点 2 笛卡尔坐标

0				
1				3
0				
1		2		
y/x	0	1	0	0

图 11.22　点 3 笛卡尔坐标

0				
1		3′		3
0				
1		2		2′
y/x	0	1	0	0

图 11.23　绑定难题

	a	b	c	d	
0					e
1		3′		3	f
0					g
1		2		2′	h
y/x	0	1	0	0	

图 11.24　合取编码

解决方案 1:合取编码(conjunctive coding)

一种解决方法是对**每一对可能的 X 和 Y** 对指派一个单元,如图11.24。因此,点 2 的节点为(b,h),它的重影 2′是节点(d,f)。点 3 与它的重影 3′的节点与之类似。这就解决了在这个简单例子中的绑定难题。但随数值的增加,很快会加重计算负荷。 321

如果我们在操作上将这种表征系统精确地定义为:点在平面移动一定标准间隔而产生的不同编码数目——那么为精确地获得某一具体点的表征,需要合取编码的数目与这个点数值的平方根成比例(参见 Rumelhart and McClelland,1986a:90)。显然,这并不是最佳方案。

解决方案 2:粗编码(coarse coding)

"粗编码"是一很有前景的技术。使用这种方法,将单元视作二维圆环,多个单元环在输入域相互交叉。使用大的单元环好,还是小的单元环好呢?凭直觉,也许会选择后者。但事实上是前者更有效,因为表征的精确度与单元环的数量(n)以及它们的半径(r)相关:a=n×r("="指"成正比")。所以增大单元环的半径也就增加了精确度(参见 Rumelhart and McClelland 1986a:91)。如果交叉点(普遍特征)在空间中的分布较为广泛,那么这种类型的表征会表现得更好。但如果单元间靠近得过于紧密,又会有被编码为是同一组单元的危险。这种风险是获得精确度必须要承担的。

微特征

我们将在第 12 章中讨论,按照斯莫琳斯基的观点,联结主义模型是介于经典概念层(例如,框架)和神经层间的操作(参见 Smolensky,1988a)。他称之为"次概念"层,也就是"直觉处理器"层。在这个层次上,"表征是基于众多单元参与的复杂激活样式,每个单元都能够参与若干种样式"(1988a:6)。样式作为整体才具有概念解释或者语义特征,并不涉及构成它们的单元:"那些单元并没有概念的语义特征:它们属于**次概念**"(1988a:6)。那么,样式在语义特征中所起到的作用是什么?目前还没有一种一般性的答案,因为每种模型都有其将激活样式与概念解释相联系的独特程序。在实际操作中也有一些总的应用策略。其中一种是借助某领域内高层次概念描述中的分类,如向节点指派语音或者图像特征。另一种是在使用多维排列、层级聚类分析等训练方法时,分析隐层单元的表征形成过程。 322

对于第一种方法，需要注意到模型适用领域的高一层次描述可以是常识性的。例如：

> **咖啡**的联结表征，是**装有咖啡的杯子**的表征减去**没有咖啡的杯子**的表征……事实上，除激活样式外余下的都属于激活特征，例如，具有扁平表面的棕色液体，具有曲线边缘和底面的棕色液体，与瓷器相关、热的、有烧焦气味的棕色液体。这就表示了**咖啡**，在某种意义上——**咖啡是在杯子里的内容**。(1988a:16)

随后，斯莫琳斯基给出了有关图解，如图 11.25 所示。

323

装有咖啡的杯子表征	
单元	微特征
●	竖直容器
●	热的液体
○	杯子接触桌面
●	具有曲线表面的瓷器
●	烧焦气味
●	瓷器内装有棕色液体
○	长方形银质器皿
●	带有握柄
●	带有曲线边缘和底部的棕色液体

空的咖啡杯子表征	
单元	微特征
●	竖直容器
○	热的液体
○	杯子接触桌面
●	具有曲线表面的瓷器
○	烧焦气味
○	瓷器内装有棕色液体
○	长方形银质器皿
●	带有握柄
●	带有曲线边缘和底部的棕色液体

咖啡表征	
单元	微特征
○	竖直容器
●	热的液体
○	杯子接触桌面
○	具有曲线表面的瓷器
●	烧焦气味
●	瓷器内装有棕色液体
○	长方形银质器皿
○	带有握柄
●	带有曲线边缘和底部的棕色液体

图 11.25 咖啡矢量(来源:Smolensky,1991b:208-9,图 2—4;经 Kluwer Academic/Plenum Publishers 许可复制)

再来看第二种方法。对隐层单元的分析说明,网络能够形成有关类别的表征,这种类别对于“处在刺激中”是有意义的,但显现得并不是非常清晰。这些类别在 NETtalk 网络的隐单元层分析中发现是一种潜在的固定排序。

至此,我们已经将联结表征分为定位式(定点)和分布式,并注意到在某些模型中对于一些信息来说是定位式(NETtalk:字母以及他们的独有特征),而对于另一些信息来说是分布式(NETtalk:单词和音素)。还存在其他一些可能的表征种类(如图 11.26):

类型 1 :有关这种类型的示例是我们刚提到的斯莫琳斯基的“咖啡”。某一样式可表征,比如,一只杯子;它的构成节点可表征有关这个概念的微特征。NETtalk 的输入层和输出层也符合这一类型。

类型 2 :有关这种类型的示例是 NETtalk 的隐单元层,它的激活样式表征元音和辅音,但所有的单个节点并没有解释。

类型 3 :有关这种类型的示例是将在第 13 章中讨论的命题网络,样式表征命题,但其构成节点并没有特殊的语义值。

类型 4 :有关这种类型的示例是递归网络。在递归网络中,表征从网络随时间变化采集调整不同结构关系的指令中获得。也就是斯莫琳斯基所说的“张量空间(tensor product)”表征。

324

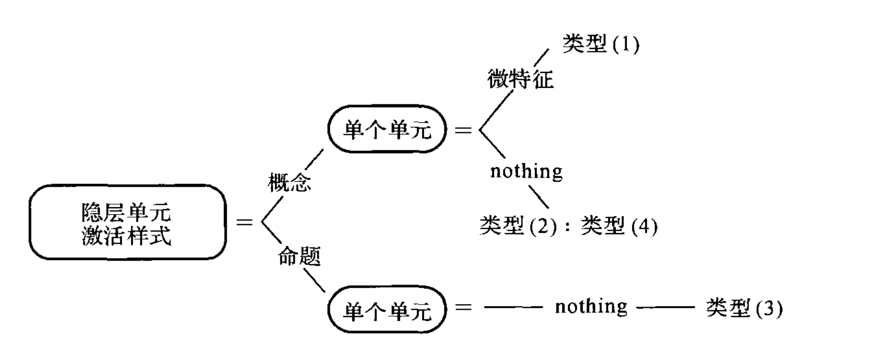


图 11.26 联结主义表征类型(来源:Ramsey,1992:261,图 8.1;经 Oxford University Press 许可复制)

采用分布式表征的一些优点

在联结主义模型中,分布式表征比之定位式表征具有很多优点,我们将在后面两章讨论。正是由于分布式表征的这些优点,才引起人们对联结主义机器的关注。我们在这里对这些优点略作了解,无疑是有益的。分布式表征被人们普遍认可的优点是:具有内容寻址性、样式完备性、(自发)归纳

能力、容错性，以及网络受到部分损坏时功能渐次衰减和较强的再学习能力等。所有这些优点彼此依赖，而非完全独立。如我们在第6章中讨论的，“内容寻址性”与定位寻址性相比较——材料的储存是依据它所表征的内容，而非(任意的)地址。如果被存储的信息(在内容上)与所期望的信息接近，那么可以直接对之提取。“样式完备性”，是指采用分布式表征的系统能够正确地识别部分缺省的输入。“自发归纳”，是指系统能够激活与目标节点相关的节点的能力，具体特征越具有“普遍”性，越能激活更多的节点。“容错性”，是指系统能够兼容错误信息，或者即使输入信息有误，也能找到正确的答案或者最适合的答案。“渐次衰减”，是指系统总会出现程度不同的损坏情况，但系统并不会崩溃，而是表现为整体在不同系统性能上，如速度或精确度，有所下降。最后，“较强的再学习能力”，是指如果系统受到破坏，如它产生了大量噪音，那么再次对其训练，可以以比先前更快的速度重新学习，甚至会提高先前训练所忽略的节点表现——也就是说，再训练时系统提高了“归纳”能力。

问题2:联结表征的本质

我们已经介绍了三种表征的承载者:输入单元、输出单元和隐层单元；325 获得表征效力的三个来源:程序员指派、学习和先天固有。程序员指派表征似乎没必要讨论，因为它只是简单地把表征本质难题退回了一步:程序员所指派的表征用什么(如何)表达其所表征？这一问题，当解释单元会去做什么时就变得清楚了，这一问题的意义就是在探明当给定模型信息时它做了什么。通过深入了解模型行为如何能被解释，可以知道赋予机器的表征所倾向于具有的内容。程序员所指派的语义，最终要么归结于模型通过学习获得，要么归结于模型先天固有。最值得关注的是通过学习获得的表征——机器自身对刺激形成的表征(或者通过无监督学习偶然发现，或者有监督学习)。明显地，输入单元和输出单元是通过程序员指派语义，而隐层单元则通过训练获得语义。

11.5 一般联结主义

我们已经概述了有关联结主义机器的一些基本概念:单元、激活、连接、权值、计算和学习。从中，我们可以发现联结网络具有的一些普遍特征和专

属特征,找出所有联结网络的共同点和区分点无疑是有益的。我们将采用这样的策略,即总结一些有关联结网络的通用论题,用于回答有关对联结网络提出的一般性问题。

十个问题

Ⅰ 静态特征

结构

单元

1. 总共有多少个单元?

连接

2. 什么是连接样式(几何、"空间"排列)?
3. 哪种连接是兴奋的,哪种连接是抑制的(如果有的话)?
4. 激活作用的传递方向是什么?

表征

5. 单个单元在每一单元层中表征什么,如果有的话? 326
6. 单元组在每一单元层中表征什么,如果有的话?

Ⅱ 动态特征

计算

7. 网络如何从输入计算其输出?

编程

单元

8. 什么是"激活传递"规则(如果有的话,包括阈值和衰减)?

连接

9. 什么是连接权值?

学习/训练

10. 使权值或者阈限变化的程序是什么?

一般联结主义机器

1. 是连接计算单元的网络。

2. 每两个连接都有某一强度或权值。
3. 每个单元具有某种传递激活值的规则。
4. 某些单元是输入单元,它们的(可能是联合起作用的)激活表征其输入。
5. 某些单元是输出单元,它们的(可能是联合起作用的)激活表征其输出。
6. 数据/信息储存在激活了的单元和连接权值之中。
7. 学习法则用来调整连接权值。
8. 计算是对激活层次和连接权值的算法操作

注释

[1] 对之更早的阐述可以在James(1980)中找到。
[2] 在一特定组群中,丢失任何两个神经元的概率是千万分之一——这是不太可能的。

【思考题】

基本概念和术语

327 单元的三个功能部分是什么?
什么是"群收"和"群发"?
工作网络的编程是指什么?
是什么"净"激活传递规则?
什么是"输入矢量"?
什么是"输出矢量"?
上面两个术语是关于什么的计算?
给定某一工作网络和输入矢量,运用(N)计算它的输出矢量。
工作网络如何调整缺省输入?
单个工作网络如何运行多于一种的联结样式?

学习和训练

学习的主要类型有哪些?
什么是赫伯学习规则(三个部分)?
给定某一工作网络、输入矢量和输出矢量,运用赫伯规则(H)计算权值分配。
给定某一工作网络、输入矢量和输出矢量,运用Delta规则(D)计算权值分配。

表征

什么是定位式表征?
什么是准分布式表征?
什么是(完全)分布式表征?
定位式表征具有的三个难题是什么?

分布式表征的优点是什么?

联结主义表征是如何关涉它们所关涉的事物的? 328

一般联结主义

一般联结主义机器的两个“静态”层面是什么?

一般联结主义机器的三个“动态”层面是什么?

工作网络如何储存现时知识?

工作网络如何储存命题知识?

【推荐读物】

概论

Schneider(1987)包含关于联结主义范式转换的介绍。Anderson(1995)是有关联结网络较好的研究导论。McClelland and Rumelhart(1988),以及 Caudill and Butler(1993)是两本清晰和有益的导论性著作,并附有联结机器的模拟磁盘。其中,在 Caudill and Butler(1993)第 1 卷中,涵盖了基本网络构成和训练的内容;在第 2 卷中,提到改进网络构成和训练的有关内容。新近出版的 Plunkett and Elman (1997),以及 McLeod et al. (1998)中也附有模拟磁盘。McClelland and Rumelhart (1988),Levine (1991),Caudill and Butler (1993),以及 McLeod et al. (1998)考察了各种联结结构,其中包含文中所讨论的问题。

基本概念

Feldman and Ballard(1982)是较早的关于联结主义的重要文献(人们普遍公认的)。迄今为止,或许是最有影响力的关于联结主义模式的导论,可见 Rumelhart and McClelland(1986)第 1 卷第 2 章。该书对并行分布加工的一般结构以及基本概念有很好的阐述。Rumelhart(1989)对与这章相关的大部分内容有很好的概括。其他对联结主义基本内容有较好研究的,可见 Wasserman (1989)第 1 章,Clark(1989)第 5 章,Churchland(1990),Quinlan(1991)第 2 章,Bechtel and Abrahamsen(1991)第 1—3 章,Haberland(1994)第 6 章,Bechtel (1994),Stillings et al. (1995)第 2.10 节,以及 Elman et al. (1996)第 2 章。虽然 Ballard(1997)技术上的内容更多些,但仍有很多对我们有所帮助的细节。对递归网络及其语言更多的讨论,见 Elman(1992)。

学习和训练

在 Rumelhart and McClelland(1986a)第 1 卷中,第 5 章是有关竞争学习的内容,较好地介绍了学习类别,并作了具有重要影响的讨论;第 8 章是有关基于误差传播学习内部表征的内容,介绍了反向传播、广义 Delta 规则,具有重要的历史地位。Hinton(1992)对主要的监督和无监督训练技术作了精练和权威性的考察。 329

相关数学

Rumelhart and McClelland(1986a)第 9 章对并行分布加工所涉及的相关线性代数内容作了深入介绍。其他相关的简短讨论，参见 Wasserman(1989)附录 B"矢量和矩阵运算"，Levine(1991)附录 2"神经网络的积分和微分方程"，以及 Caudill and Butler(1993)第 1 卷附录 D。

表征

最早最有影响力的对分布式表征的介绍，可见 Rumelhart and McClelland (1986a)第 3 章"分布式表征"。对于表征的神经生物基础的讨论，参见 Nadel et al.(1986)和 Feldman(1989)。Sterelny(1990)第 8 章对联结主义表征作了哲学讨论。更多的哲学视角的探讨参见 Cussins(1990)。Smolensky(1990)是个简短而又具有权威性的总体评论。van Gelder(1991)，Goschke and Koppelberg (1991)，以及 Ramsey(1992)中对联结主义表征作了批评性的评论。

文献选集

关于联结主义，具有代表性的文献选集主要有 Nadel et al.(1989)，Morris (1989)，Horgan and tienson(1991)，Ramsey，Stich，and Rumelhart(1991)，Dinsmore(1992)，Davis(1992)，Clark and Lutz(1995)，以及 MacDonald and MacDonald(1995)。

12 心智的联结计算理论

12.1 引 言

在第二部分，我们通过对早期心智表征理论(RTM)加入计算限制，得到 330
了一般心智计算理论(CTM)，如下：

(CTM)

1. 认知状态是具有内容的**心理表征的计算关系**。
2. 认知过程(认知状态的转换)是**心理表征的计算操作**。

我们还指出，通过对 CTM 进一步增加具体的数字限制，就可以得到心智**数字**计算理论(DCTM)。有两点需要考虑：第一，具体的数字**结构**(存储和控制)；第二，具体的数字**表征**。我们现在开始讨论与 DCTM 类似的联结主义结构，并探讨联结主义与心智理论的关系。

12.2 心智的联结计算理论

早期获得**联结**计算理论的方法，是通过对 CTM 进一步添加具体的联结主义限制。那么，对于 CCTM 也显然要思考它的结构和表征——计算的结构和表征是联结的。DCTM 与 CCTM 在操作方式上有很多不同(它们甚至使用不同的数学方法)，但 CCTM 与 DCTM 也有很多**相似**，下面这些特征也是 DCTM 所具有的：

331 1. CCTM 模型具有**符号**：节点和节点的矢量。

2. CCTM 模型是程序化的：通过设定激活传递规则，连接权值和阈值(thresholds)而获得。

3. **当下数据**：激活节点的矢量。

4. **已处理数据**：存储在连接权值中。

5. **计算**是矢量间的转换。

这就使我们期待，心智联结**计算**理论非常具有可行性。还能得到更多吗——它是否具有实践意义？似乎看来还没有。许多观察者都同意泰森(Tienson)的评论："联结主义的认知观念，作为**一种心智理论**时，并没有增加我们任何关于心智的了解，更不用说它的**具体现实机器**了……[它]只可能会告诉我们，什么样的网络活动构成了心智活动，以及为什么。但到目前为止，联结主义还没有告诉我们，如何回答这些问题"(Tienson, 1992: 2)。著名联结机器设计者斯莫琳斯基(Smolensky)认为，联结主义作为一种认知模型的框架，它的进步有点像亚里士多德信徒，"我们已经具有了引人关注的技术和有建设性的提议……以这种新的计算取向，在智能史上，我们正在以某种方式接近亚里士多德的立场。如果任何联结主义的热情拥护者认为，我们能够以真正的认知模型取得这样的位置，那么恐怕他们就错了"(Smolensky, 1991b: 287)。

然而，运用 CTM 模型，我们可以对基本的心智联结计算理论初步版本的主要观点，形式化地阐释如下：

(B-CCTM)

1. 认知状态是具有内容的**心理表征**的**计算关系**。

2. 认知过程(认知状态的转换)是**心理表征**的**计算操作**。

3. 计算的结构和表征(1 和 2 提及的)必须是**联结的**。

在介绍 DCTM 时，我们尝试从它与心一身关系等其他问题的关系上，对其进行说明。我们粗略地限定了这个理论对于认知状态和过程的范围，然后根据 DCTM(所谓的"命题态度")对一些认知状态具有什么特征进行了阐述。心智联结计算理论的发展比先前的数字计算理论慢得多。目前仍有许多问题未解决。例如，有人提出[1]，CCTM 是否对心一身关系难题提出了与其他心智理论明显不同的观点，仍存在争议。思维语言假说在多大程度上 332 可以转换为 CCTM 也存在争议。如果将思维语言的观点扩展至联结主义模型，那么我们能得到联结主义与第 8 章(DCTM)类似的表述：

(CCTM)

1. 认知状态是具有内容的**计算心理表征**(在思维语言中)的**计算关系**。

2. 认知过程(认知状态的装换)是那些**计算心理表征**(在思维语言中)的**计算操作**。

3. 计算的结构和表征(1 和 2 提及的)必须是**联结的**。

现在,我们来探讨 CCTM 的一些理据以及它的典型特征。

12.3 CCTM 的理据

溯寻 CCTM 模型的理据,发现主要源于两个方面:它的网络加工活动与人类行为相似,它的网络加工特征与人脑相似。

人类行为

回顾我们之前讨论的 Jets & Sharks 网络和 NETtalk,这些模型都具有一些类似人类行为的特征:

Jets & Sharks 网络

这种模型可以:

1. 获取特定个体("原型")的具体特征。

2. 从各个特定个体("原型")的存储知识中,抽取这一组对象的主要倾向。

3. 填充模糊的缺省值。

NETtalk

1. 比之音位,能够更快速、准确地获取重音。

333

2. 按照"幂定律"进行学习(如,学习率和学习误差在双对数标度上函数几乎是一条直线)。

3. 间断学习(spaced practice)(新旧词汇交替出现)比连续学习有更高的效率(从艾宾浩斯开始,人们就知道这种提高人类学习绩效的技巧)。

4. 能够逐渐提高对新词汇的相对可靠的发音,NETtalk 学习的内容越多,它的行为表现得越好。

5. 当权值随机受到损害时,它的整体表现也将逐渐地("逐步地")衰减。

6. 损害后,再次学习先前的学习项目,比第一次学习得快。

7. 基于层级聚类分析（发现所有字母的读音，在隐层单元的发音活动中的不同特点），发现 NETtalk 将元音矢量聚集在一起，区分于辅音。另外，在元音和辅音的分类中，读音类似的元音（以及部分读音相似的辅音）也聚集在一起。

大　脑

支持 CCTM 模型的另一个原因，是强调 CCTM 模型与人脑有很多相似之处。理由如下：

1a. 心智/脑模型，不应该将两者完全等同，更准确的说法是心智模型和唯一的脑模型，或者说

1b. 比较好的心智模型应当与脑模型非常接近或匹配。

2. 心智的 CCTM 模型就是这样的模型，它能够较好地模拟脑的总体结构和功能。

3. 所以，CCTM 是首选的心智模型。

DCTM 认为，我们能够完全独立于人脑（硬件），对人的心智（软件）进行研究和理解，因此 CCTM 的这种观点与 DCTM 非常不同。CCTM 倾向于弄清心理现象与神经现象之间的密切关系，尽管联结主义模型并不是神经模型，但能够模拟人脑的总体结构和功能特点。CCTM 反对目前的 DCTM 在硬件上不能很好地反映人脑的总体结构和功能，因而可能会认为 CCTM 并不是一种严格意义上的以认知为基础的脑模型。如果接受了这一假设，那么就要继续考虑下面的问题。

334

大脑：结构

CCTM 模型与大脑有许多结构上的相似之处：

1. CCTM 模型中的单元类似于大脑的神经元。

2. CCTM 模型中的连接（和权值）类似于神经元的轴突、树突以及突触。

3. 脑和某些 CCTM 模型都是以层级的方式进行组织的。

4. 脑的学习可能是以调节突触的连接强度进行的，某些 CCTM 模型也这么做。

5. 脑的各个部分表现出并行兴奋和抑制的特点，某些 CCTM 模型也是这样。

大脑:功能

CCTM 模型与大脑在功能上有很多相似之处:

(1)*巨量并行加工*(massively parallel)　与 CCTM 模型一样,脑的加工过程可能都是以大量并行的方式进行的。至于为什么是这样进行加工的,一种解释是:脑的神经元约以 1/1000 秒的速度传递信号(进行一次触发),而计算机约为 1/100000000 秒——比脑快百万倍。因而,如果脑是以串行程序进行加工的,那么它在 1/10 秒内——执行视觉识别和语言理解等基础操作的时间,只能运行 100 个指令。但对于这些任务的操作不可能只需要 100 步就可以完成。鲁梅尔哈特(Rumelhart)和麦克莱兰德(McClelland)作了如下解释:"现代串行计算机是以毫微秒为单位对基础操作进行测量的,而对神经元操作速度的测量单位有时是毫秒——或者 10 倍毫秒。因此,脑的基础硬件比串行计算机要慢约 10^6 倍。然而,需要注意的事实是,我们只需要几百毫秒就能够完成非常复杂的加工。毫无疑问,感知加工、大量的记忆提取、大部分语言加工、直觉推理以及其他的加工,都能够在这个时间范围内完成,这意味着人脑对这些任务的加工必定要少于 100 步,或者说并不是串行加工,也就是费尔德曼(Feldman, 1985)所说的**程序指令的 100 步限制**(100-step program constraint)。另外,还需要注意的是,单个神经元可能并不能计算非常复杂的函数。单个神经元计算的函数,应该不会比数字计算机中的单个指令要复杂"(1986a, vol. 1:130-1)。这样的限制通常称之为"100 步法则"。　335

(2)*内容寻址性*　与 CCTM 模型一样,脑使用信息(内容)的片段就能获得信息的全貌——以内容-(content-),非定址-(not location-)的方式进行寻址。

(3)*分布式存储*　与 CCTM 模型一样,脑的每一种记忆似乎都没有某种特定的存储地址,而是将信息分布于脑的众多区域之中,每个区域都可以参与储存信息的若干片段(回顾第 3 章对拉什利的评述)。对于某种单个记忆,有人估计大约有 700000—7000000 个神经元参与存储这个记忆的痕迹,也有人估计是有 1000 个神经元参与。我们已经了解到,在 NETtalk 中约有 20% 的隐层单元参与单个记忆的编码。

(4)*渐次衰减*(graceful degradation)　与 CCTM 模型受到损坏一样,如果脑受到了损伤,会递级产生行为缺陷——也许不会出现明显的错误,但是系统的整体功能却不能很好地运行。

(5)*对于杂乱和缺省输入的不敏感性*　与 CCTM 模型一样,脑似乎能够轻松地加工杂乱和缺省输入——回顾当玫瑰网络的输入是缺省的时候,它

的行为表现。

在后面的第 13 章中，我们会发现 CCTM 的这些功能优点其实并非是没有争议的。

12.4 联结主义与联想主义的历史关联

我们已经介绍了联结主义机器的主要组成部分，以及它们是如何进行组织、计算和训练的。正如我们前面所指出的，这种近来迅速兴起的模型的思想发展脉络，可以向前追溯很多个世纪，包括经典联想主义（第 1 章）、巴甫洛夫与条件反射（第 2 章），拉什利与赫伯（第 3 章），感知器（第 4 章），以及鬼蜮模型（第 6 章）。现在我们更详细地探讨联结主义和联想主义的关系。

一定会有人对早期的联想主义工作网络与部分联结主义的工作网络在几何上的相似性，感到惊讶。这里，我们再次将詹姆斯（1980）（见图 1.6(a)）与鲁梅尔哈特和麦克莱兰德（1986a；见图 12.1）的整体回忆模型进行比较：

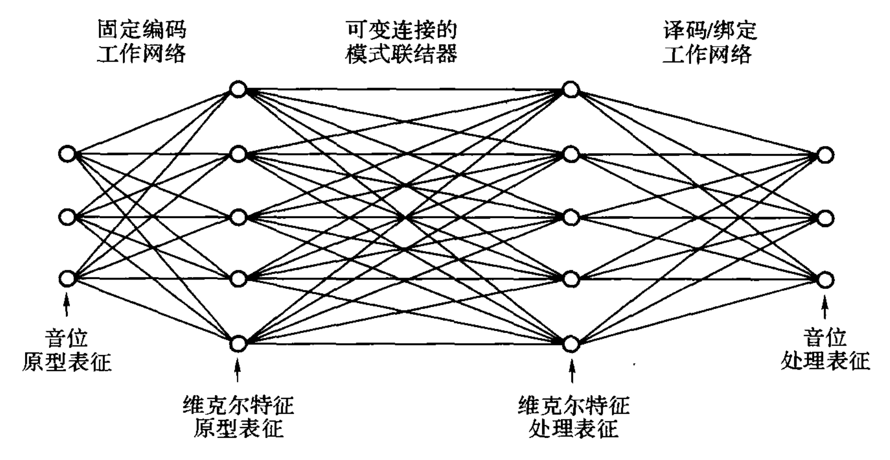


图 12.1　时态工作网络（来源：Rumelhart and McClelland，1986，vol. 2：222，图 1；经 MIT Press 许可复制）

336　它们之间有哪些一般性的关系呢？如果将第 1 章的“经典联想主义”与第 11 章的“经典联结主义”相比较，会发现前者似乎是后者的一种特例，从这个意义上说，完全能够设计一种联结主义工作网络，使其具有表征与联想能力。例如，可以用输入节点作为感知加工的感知传感器（sensory transducers）（思考 NETtalk 中的字母探测器），也可以使隐层单元能够对观念进行编码，还可以使输出单元产生行为（思考 NETtalk 的发音特征）。联

结主义模型，还可以通过设定它的连接权值，使其获得感知间，观念间，或者感知和观念之间的相似特征(或对比特征)，因为这种联想(连接权值)，能够记录感知间等的在空间和(工作网络正在运行的)时间中的连续性。也可以认为，因感知间等的相似特征，而在 NETtalk 中产生隐层单元束，就是一种联想。相似特征与单元束集之间的因果关系是很难处理的，但休谟在他的理论中涉及了这一点，而且只有在分析这个问题时，他才使用了邻近和连续的概念。

除了联结主义具有普遍性和灵活性之外，联结主义与联想主义之间还有什么不同吗？柏克德(Bechtel)认为："对于英国的联想主义者，联想的最普遍基础是在形成观念的过程中观念，或空间，或时间的相似性，这些工作网络的这种最基本特征使得它能够直接与感知觉相连接……因此与 PDP 理论者提出的联结工作网络是有很大不同的。在联结主义的工作网络中并没有直接与感知输入或运动输出相连接的单元，只是有很多单元间接地与它们相连接"(Bechtel，1985：57)。柏克德的观察可能只是发现两者之间的程度不同而已，对他来说，联结主义关注于输入和输出中间的中心加工部分，而联想主义却继续将理论扩展至周围神经。但是他没有从文献中给出这种不同的例子，也很难说明为什么原则上将联结主义模型扩展到周围神经是不可能的。 337

似乎某些联结主义模型和联想主义工作网络至少存在两点不同。第一，在联想主义工作网络中，获得联想的是观念，联想的建立是通过增强观念之间的通路(连接)。在**局部**(local)联结主义工作网络中，如 Jets & Sharks 网络，运用的是模拟的方式，联结主义工作网络中的一个**节点**模拟一种观念，节点间的连接权值模拟观念间的联想，并且它用于学习的联结原则主要是修改连接的权值。但在**分布**联结主义工作网络中，观念的模拟是用节点集合或节点矢量，并且矢量间不存在连接以及可变的连接权值——连接只存在于组成矢量的节点间。换句话说，并不存在从一个矢量到另一个矢量的变换原则，可以用于模拟系统中从一种观念到另一种观念的联想。第二，一般而言，对于联想主义，观念是感知觉的**复制物**，正如我们看到的，这也是联想主义认识论的主要特征。但是，因为联结主义的连接权值不能等于 1，所以信息也就不是在输入单元和隐层单元间的复制。

12.5 联结主义的解释：PTC

我们首先需要强调，联结主义对认知解释的形式化表述，以及它的哲学

寓意仍未达成一致意见。麦克劳克林(McLaughlin)对这一现状作了很好的总结:“一部分人认为联结主义并不属于常识心理学(folk psychology),而另一部分人却仍在努力维护联结主义属于常识心理学;一部分人认为联结主义属于一种准工具取向,用于探究意向的归属性(intentional ascriptions),另一部分人却认为联结主义与意向实在论(intentional realism)相吻合;一部分人认为,联结主义可以使人们放弃认知过程是在句法结构上对心理表征进行操作的观点,另一部分人却声称,联结主义提供了一种最有前景的对表征进行认知操作的解释;一部分人认为联结主义与联想主义的无意识加工相融,另一部分人却认为联结主义是一种笛卡尔式的观点,即认为它支持认知状态就是意识,并且隐含着认知科学应当回避无意识加工的期望”(McLaughlin, 1993: 164)。这里,我们并不是要激励所有人达成一致意见而开列这样一份清单,而是希望能从所有这些可能中,用心理学、神经学、哲学的术语,找出一种全面的、连贯的联结主义的形式化解释。斯莫琳斯基(Smolensky, 1988a, 1989)为这种解释提出了一个框架,称之为“联结主义的恰当定位”(proper treatment of connectionism, PTC)。

338

层次分解

人类的认知能力可以分解为三个最简层次,分别是概念层、次概念层和神经层。

概念层

第一个层次是“**概念**”(或“**符号**”)层。在这一层次上,认知结构可被分解为我们熟悉的概念单元,通过词汇或数据结构而获得认知结构,正如我们在“SHRDLU”中所看到的(见第5章):“金字塔”“放置”“盒”“内”(“put”“pyramid” “in” “box”)。这样的符号结构具有传统的语义解释,并且能够借助符号的形式或形状通过符号处理技术进行操作。

次概念层

第二个层次是**次概念**(或**次符号**)层(subconceptual level),是概念层描述的构成成分。次概念(也称为微特征)对应于单元(节点),而概念对应于集合、模式或节点矢量。既然联结主义系统中对认知的分类是通过模式或矢量的激活实现的,所以就存在两种类别的构成关系:模式—次模式的关系(子集)和模式—节点的关系(元素)。第一种属于概念—概念关系,而第二

种则属于概念一微特征关系。它们虽然都反映了“部分一整体”的关系，但却有所不同，如图 12.2 所示（这里忽略了一个问题：概念和次概念/微特征区分的标准是什么？）。

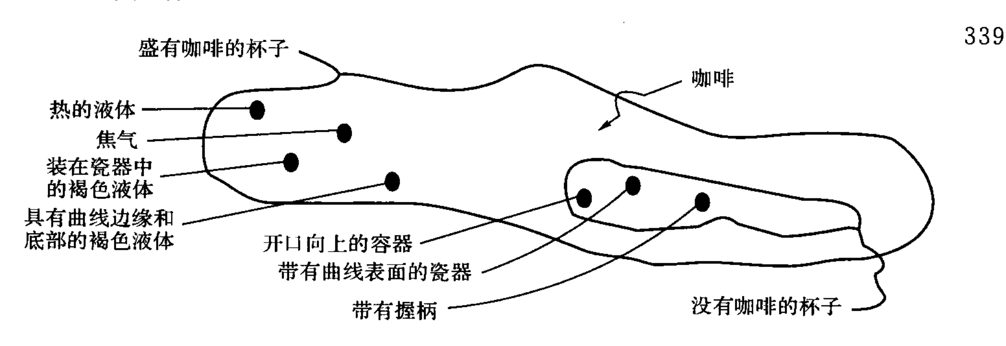


图 12.2 联结主义概念的两种构成类型

神经层

第三层是神经层。这个层次构成神经科学所研究的神经系统的结构和操作。

知识与信息的加工方式

在认知功能中，有两种重要的不同类型的处理方式：知识与信息的加工。

有意识的应用规则

第一类知识是有意识地对规则进行应用[2]。这里阐释的规则，是指在一些“语言”系统（自然语言、程序语言、逻辑和数学的符号标识法）中，应用于概念层任务域中的规则，例如科学和法律。一般而言，这类知识是社会实践和建制活动的产物，但这并不是它的标识性认知特征的本质。对于有意识的应用规则，典型的例子是一位初学者让自己重复有意识地做这样的事情：旋转、迈步、摇摆（学习打网球），或是：“i”在“e”前而非在“c”后（学习单词拼写）。

直觉知识

第二类知识是直觉知识——如感知、（母语）语言理解和熟练的动作。这类知识是以明显迅速灵活的方式习得，而不能依靠内省通达（回顾福多的“模块”或“输入系统”）。

显而易见，作为一种认知功能，**有意识地应用规则**是在**概念层**中习得，

而**直觉加工**是在**次概念层**习得。同时，我们也应该看到，这两种功能都能在**神经层**上实现或具体化。

联结主义模型层次

340　现在我们要提出的问题是，就上面的那些区分，把联结主义模型放置在什么位置才是恰当的。再来看第一种区分，根据斯莫琳斯基的观点，PTC 认为联结主义处于神经模型和传统符号模型的中间位置——涉及次符号模型（本章后文带有数字标识的引文均引自 Smolensky，1988a 原文）：

> **(11)**
>
> 次符号程式的基础层次，或次概念层，位于神经层和概念层之间。

对该层次的分析比神经学更为抽象（尽管它关注了神经结构的一般特征），但与传统符号模型相比，则更强调认知的动力特征。这与量子力学比经典力学更强调物理粒子的动力一样；经典力学只为现象提供一种非精确且仅是近似的解释，而量子理论的解释却足够精细且准确。下面让我们再来详细地分析这三个层次。

符号（概念）程式

概念层包含有意识的规则解释器，属于**符号程式**的自然域。符号程式与“文化”知识相关，如科学、法律，是在自然和科学的语言中形成的，可以将明确的指涉规则应用于“文化”知识。有效步骤理论、图灵机以及冯·诺依曼计算机程序，为我们提供了心智如何加工这些知识和执行这些指令的模型：

> **(3)**
>
> (a)在语言中形成的规则，能够为文化知识提供一种有效的形式规则(effective formalizatio)；
>
> (b)有意识的规则应用，通过一种虚拟机制，对这些规则进行序列解释，这种虚拟机制称之为有意识的规则解释器；
>
> (c)这些规则是在用于有意识的描述任务域的语言中形成的——在概念层中形成(1988a:4-5)。

341　符号程式认为：

> **(4a, b)**
>
> (a)运行在直觉处理器(intuitive processor)中的程序包含语言的形式规则，这些规则能够获得序列性的解释；

(b)运行在直觉处理器中的程序是由其元素构成的,即符号,主要用于指涉有意识地对任务域进行概念化的概念。(1988a:5)

例如,回顾 SHRDLU(见第5章)的例子,我们可以发现数据和指令能够依据常识语言获得形式化,如 PYRAMID 和 MOVE。可以概括为:

(4)

无意识规则解释假设:运行在直觉处理器中的程序,具有句法和语义的特征,与在有意识的规则解释器中运行的程序非常相似。

符号程式的**计算方式**,具有**严密**和**序列性**的特征。在符号程式中具有:

(24)

(a)离散记忆的定位性,它的记忆项目不能相互作用;

(b)对离散记忆存储和提取的操作过程,记忆中的任何一个完整项目的存储和提取,可以发生在单个最简的(基础的)操作中;

(c)对离散学习的操作过程,按照有一或一无的方式获取新的规则;

(d)对离散学习的操作过程,按照有一或一无的方式得出结论。

(e)离散的类属性,它的项目要么属于这个类别,要么不属于。

(f)离散的产生式规则,其条件要么满足,要么不满足,由条件产生的动作要么执行,要么不执行。

在符号程式中,以上的认知层次与计算机系统的层次相类似。但在计算机系统中,确切地说明符号层如何在神经层上实现并不是符号程式的一部分(见图 12.3)。然而,斯莫琳斯基反对这种观点,至少有如下理由:

(5)

(a)依据假设(4)构建的人工智能系统,事实上似乎过于勉强、缺乏灵活性,因此并不能作为真实人类技能的模型。 342

(b)依据规则熟练地表达专业领域知识的过程,似乎对于很多重要的认知领域(如常识性知识)不切实际。

(c)假设(4)没有对知识如何在脑中获得表征作出实质解释(1988:5)。

次符号程式与符号程式

斯莫琳斯基关注的是直觉加工,例如"感知、熟练的动作、流利的语言行为……总之,是所有已习惯了的熟练行为"。联结主义模型能够在次符号程式上,对这些过程作出详细和确切的描述,而符号模型对此却无能为力。这使斯莫琳斯基认为,直觉加工的实现基础不是符号程式,而是次符号程式:

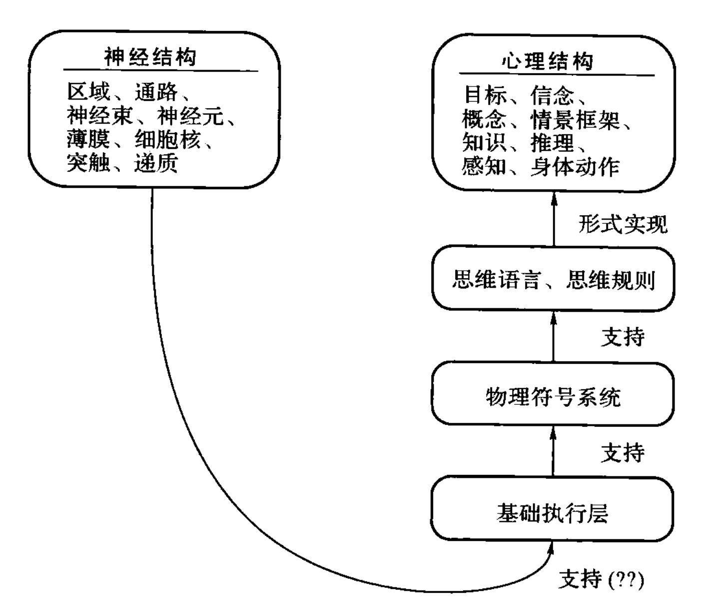


图 12.3 符号程式中的神经与心理结构(来源:Haugeland, 1997: 235, 图 9.1; 经 MIT Press 许可复制)

(7)

直觉处理器具有联结主义的某些结构(这些结构是对神经网络的一些最普遍特征进行的抽象模拟)。(1988a: 6)

343 斯莫琳斯基对直觉处理器的几种假设进行了对比。首先,(4a)对比:

(8a)

联结主义动力系统假设:在任意时刻,直觉处理器的状态可以通过数值(每个单元具有一个数值)矢量精确定义。直觉处理器的动力特征由一个微分方程控制,在这个方程中的各种数值参数构成了处理器的程序或知识。在学习系统中,这些参数依据另一个微分方程发生变化。(1988a:6)

第二,(4b)对比:

(8b)

次概念单元假设:直觉处理器具有任务域的语义特征,这种语义特征属于任务域中的意识概念。直觉处理器本质上是指由大量单元共同产生的复杂活动样式,每个单元都会对很多这种样式的产生发挥作用。(1988:6)

最后得出：

(8c)

次概念层次假设：通常只在次概念层，而不是在概念层，对直觉加工器完整的、正式的、精确的描述才易于处理。(1988a:6-7)

以上所述可以总结为(8)，是"次符号程式的基础"(1988a:7)：

(8)

次符号假设：直觉处理器是次概念的联结主义动力系统，这一系统认为在概念层上不可能对直觉处理器进行完整的、正式的、精确的描述。(1988a:7)

斯莫琳斯基指出(2.4节)，这一特征使得次概念模型不能够实现概念模型——如果能实现，则都需要具有完整的、正式的、精确的特征。

最后，斯莫琳斯基阐述了(24)符号程式与次符号程式的**计算方式**的不同，次符号程式的计算是**数据统计和并行的**：

(25)

(a)次符号计算的知识，作为大的软(数据统计)约束集合形式化； 344

(b)具有软约束的推理基本上是一种并行加工；

(c)具有软约束的推理基本上是非单调性的[3]；

(d)确定是不是一个次符号系统可用其是否采用统计推理(statistical inference)进行识别。

总之，**符号**程式的约束是不连续的、硬性的，它的推理是逻辑的、序列的，而次符号程式的约束则是连续的、软性的，它的推理是统计的、并行的。

次符号与神经

斯莫琳斯基反对将联结主义模型等同于神经模型：

(6)

神经结构假说：直觉处理器完成一件特殊的任务，它所使用的结构与脑完成这件任务所使用的结构相同。(1988a: 5)

之所以得出这个结论，是因为脑皮层与联结主义系统具有宽泛的对应关系(见图12.4)。需要注意的是，图中之所以标有"负"的对应关系，是因为有些模型不能在硬件上直接实现，而另一些则涉及特征选择的问题，需要进一步在神经学中得到确认。还有一些进一步的争论，斯莫琳斯基并没有在

这个图中列出，例如负权值难题和反向传播的生理机制难题，我们留在第 13 章中讨论。

345

脑皮层		联结主义动力系统
状态由连续变化数值 （电位、突触面积……）确定	+	状态由连续变化数值 （激活、连接权值）确定
状态随时间连续变化	+	状态随时间连续变化
神经元间的联结作用参数可变； 知识存储位置可变	+	单元间的联结作用参数可变； 知识存储位置可变
巨量状态变化	+	大量状态变化
相互作用高度复杂 （高度非同质作用）	+	相互作用高度复杂 （高度非同质作用）
神经元占据 2+1−d 的空间	−	单元不占据空间
与周围神经元具有 紧密的连接关系	−	周围单元排布均匀
与远距离神经元具有 几何映射连接关系	−	远距离单元间具有 连接关系
突触占据 3−d 空间	−	连接不占据空间
它的位置对于信号的接收 和发送具有强烈影响		
区域间的远距离投射关系 具有复杂拓扑特征	−	节点池域间的远距离投射关系 具有简单拓扑特征
远距离的相互作用 通过离散信号调节	−	所有的相互作用 通过非离散方式调节
在单个神经元内存在复杂的 信号整合机制	−	信号整合是线性的
具有无数种信号类型	−	只有一种信号类型

图 12.4　神经与次符号层（来源：Smolensky，1988a：9，表 1；经 Cambridge University Press 许可复制）

模型与程式的关系

根据“PTC”（联结主义的恰当定位），脑的结构及其功能的特征（当然是在原则上）可在神经层次上得到精确描述。直觉处理器的特征（同样也是在原则上）可由联结主义模型在次概念层次上，即通过节点层、节点矢量、连接权值，进行精确描述。但是，有意识的规则应用与概念层的一般关系是怎样

的呢？对于这个问题，人们的认识更加模糊，也具有更多的争议。一方面，正如我们前面指出的，节点矢量用于构成概念，如由节点组成的具有次概念解释的“杯子”概念。所以，联结主义工作网络似乎是在更高层次（节点矢量）上，能够给出（原则上）概念层认知——我们真实的认知活动，确切的描述。另一方面，我们的整个概念层组织器官，还具有自然语言、数理符号运算等能力。具有这些系统的概念层的活动，似乎可以由运行在如 vNMs 或 PSs 等传统的符号结构上的人工智能的传统符号表征图式（见第 7 章），得到精确描述。但必须强调一些事项：这两种模型的计算方式并不能相兼容，传统符号模型遵循（24）（见上文），而联结主义模型遵循（25）。 346

两种模型的综合理论

一种较吸引人的建议也许可称之为“综合理论（hybrid theory）”——这种理论用符号程式解释有意识的规则应用，同时用次符号程式解释直觉加工——两者都没有错误。但是斯莫琳斯基（1988a，第 6 部分）指出，这种建议是存在问题的。例如：（1）这两种理论是如何交流的？（2）综合系统是如何依据经验而演变——从有意识的规则应用转变到直觉？（3）综合系统如何阐释实际人类规则应用的易错性？（4）综合系统有助于我们更进一步理解有意识的规则应用如何在神经上实现吗？

PTC

鉴于这些问题，以及符号程式无法在次符号程式上实现，斯莫琳斯基选择了对立面，将次符号层作为结构基础，而符号层可看作是与次符号层近似的，实质上是一种派生物。PTC 的解决方法是，用自然语言或传统“符号”表征图式得出的高层级描述，只是一种**近似的描述**——只能近似地描述人类真实的认知活动。这就如同经典宏观物理学，只能对微观物理学所精确描述的内容，给予粗略的近似描述一样，见图 12.5。

近似执行

确切地说，“近似”是一种什么关系呢？在多数情况中，我们并不是很清楚，但是斯莫琳斯基提供了一个特殊的例子——能够直觉加工语言的联结主义系统，如何实现（近似地）一个产生式系统。在斯莫琳斯基（1988a，第 6 部分）中说明了这种情形是如何进行的。首先，语言能力是直觉知识，所以我们假设联结主义系统具有语言能力。然后，这种能力可以用于编码语言表达，作为活动样式。所有的活动样式都能够存储于联结主义的记忆中。 347

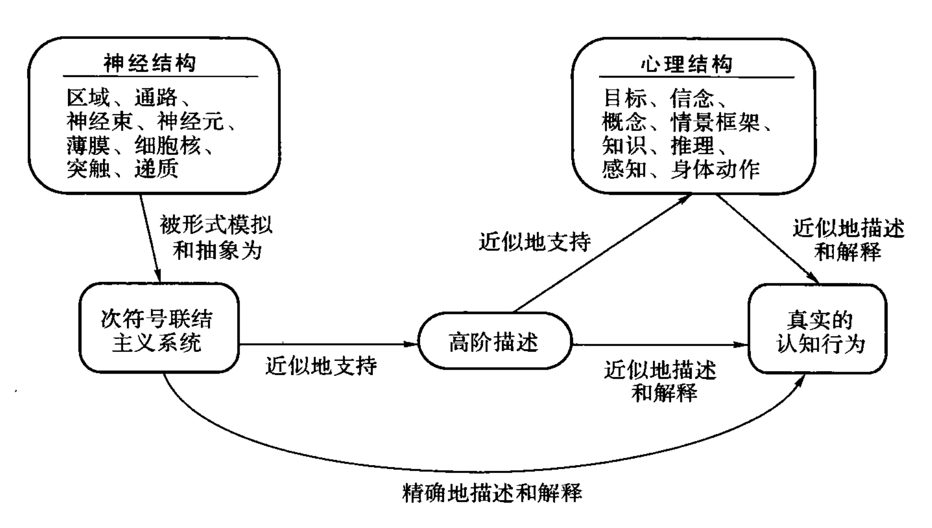


图 12.5　次符号程式中的神经与心理结构(来源:Haugeland, 1997: 237, 图 9.2; 经 MIT Press 许可复制)

因为这些表达中的一部分可以作为规则,从而我们可以使用这些规则解决问题。例如,我们能够在联结主义记忆中存储一组产生形式:

A(条件)→B(行为)

这样,给定输入 A,联结主义完备样式就能回忆起整个产生形式。通过操作产生式规则得到了 B,然后作为联结主义系统的输出,执行 B。斯莫琳斯基总结道:

(16)

在自然语言中,能够表征和处理语言的结构,是人类直觉处理器的一种能力;次符号程式认为,这种能力可以在次概念的联结主义动力系统中模拟。通过将这样的语言能力与联结主义系统的记忆功能结合在一起,序列的规则解释就能实现。(1988a:12)

当然,这里还有很多东西需要说明,也有很多疑问需要解答:

348

(Q1)符号层、次符号层次与符号、次符号程式之间的关系是什么?

(Q2)符号程式与神经层之间的关系是什么?

(Q3)有意识的规则运用与符号、次符号层次及程式之间的关系是什么?

(Q4)直觉加工与符号、次符号层次及程式之间的关系是什么?

探究这些问题会使我们脱离现在想要阐述的内容,所以就此保留。

意识、认知和内容

现在我们已经获得了(理想的)联结主义模型的整体图景,既是一种在联结主义概念层(节点矢量)上实现有意识的规则应用的具体模型,也是一种在联结主义次概念层(节点)上实现直觉加工的具体模型。但是,我们仍未触及到如何使一种工作网络(或脑)状态成为一种意识状态或者一种认知状态的一般性解释。因为认知状态是表征的,所以我们也需要对表征状态的特征(及对表征的学习)给出一种解释。

意识

根据 PTC,意识的必要但非充分条件如下:

(17)

意识:意识的内容仅仅反映了活动样式的大尺度结构:活动的次样式可以延伸到工作网络的大范围空间区域,并且能够在相对长的时间内保持稳定。(1988a:17)

显然,这种对意识的空间和时间的定义相当含糊,而且只有一个必要条件也不能令人满意。我们想知道的是:究竟工作网络的什么特征使得这些大范围的、稳定的样式成为意识的内容——回顾意识的"难问题"(见第 9 章)。

认知

因为次符号的基本原则,既不是概念层也不是神经层上的内容,那么,在何种意义上,这些模型是**认知**原理的具体化而不是神经科学原理的具体化呢?是什么将这些动力的认知系统与非认知系统区分开来的呢?根据 PTC:

349

(19)

认知系统:说一个动力系统是认知的,它的必要条件是,它能够在各种复杂多变的环境条件中,找到数量众多的目标条件。目标的组成部分以及可容纳的环境条件的变化种类越多,系统的认知能力越强。(1988a:15)

斯莫琳斯基说,"只有动力系统的复杂性"才能将认知系统与恒温器或河流区分开来。这就引起了许多重要疑问:只有一个必要条件的(19),究竟是想告诉我们什么呢?如果认知存在一个程度问题——那么恒温器有低水

平的认知吗？又或者，如果认知不是程度问题，它要么出现，要么不出现，那么，认知的**能力高低**何以依随复杂程度而发生变化呢？同时，我们暂且同意这些系统是“认知”的，那么这种“认知”与前理论概念“心理”的关系是什么？

通过复杂多变的环境条件获得目标满足，这种过程与认知相联系。这种观点至少可以追溯到纽厄尔和西蒙(Newell and Simon，1979)关于物理符号系统的讨论(可参见本书结语)。与斯莫琳斯基的区别在于，纽厄尔和西蒙明确提到的只有“智能”，不是认知：“我们测量这个系统的智能，是以当它面对繁杂变化的、困难的、复杂的任务环境时，它是否有能力到达目标状态。”如我们在讨论图灵测试(第9章)时指出的，一般性言语对于我们描述机器的“智能”特征有更大的自由度，并且描述它们的行为也会更具体，说机器具有“智能”当然比说机器具有“思想”或“认知”更易于说明。

认知的突现

次符号层与神经层具有怎样的关系，才能够生成或者确定特殊意义的认知或一般的心理状态？对于这一点，斯莫琳斯基诱人地指出，“联结主义模型有可能为我们提供揭开已经持续了几千年的心－身难题最重要的一步”(1988a：3)，但是他没有进一步详细论述这一观点。联结主义研究学派的PDP(并行分布加工)观点是，认知(以及一般的心理状态)因神经构成要素的交互作用而突现，新的突现层包含很多独特的特征，需要一些新的概念和词汇来描述它们。理解这个突现的认知层需要理解它如何由突现实现：“理解不同层次和组织结构如何突现这些新的和有用的概念。只是试图理解认知的本质特征，是如何在工作网络中因连接单元的交互作用而突现的。(350)我们确信突现的现象的存在，意味着**不可能通过孤立的研究低层次元素而获得对这类现象的理解和描述**……整体不是部分之和。因为整体的各个部分之间存在非线性的交互作用。但是，这并不是说低层次元素的属性与高层次的组织无关——恰恰相反，我们认为**为了获得对高层次组织的理解，主要还是需要对低层次单元间交互作用的研究**”(Rumelhart and McClelland，1986a：128)。显然，宣称认知突现于低层次的交互作用，并没有告知我们这一“突现”关系究竟是什么。在何种意义上，虽然这些新的、高层次的、突现的现象不可仅从低层次的事实获得理解或描述，然而他们还是认为首先需要理解低层次的现象？此外，对于身心关系“突现”如何与其他意见，如同一论或随附论等进行比较呢？这些问题都亟待进一步研究。

语义内容

次符号系统的状态是如何获取它们的内容——它们的意义和真值条件

的呢？根据 PTC：

(22)

次符号语义：认知系统容纳了在各种环境条件下的各种内部状态。在一定程度上，认知系统在各种环境条件下寻找满足它的目标条件，就特定的目标条件而言，系统的内部状态就是相应环境状态的**真实表征**(veridical representations)。(1988a:15)

事实上，次符号系统能够生成对于环境的真实表征，是通过从环境中抽取信息，然后通过学习程序将信息内在地编码为权值的结果。

SCDS 和内容

概念层认知结构在哪里获得它们的语义解释？语义解释又如何在联结主义系统中组织？在一些系统中，如 Jets & Sharks 网络和 NETtalk(见第 10 章)，我们已经看到，语义特征通常由工作网络的设计者指派给输入层和输出层，然后经过训练，再由隐单元层(如果具有)形成一些语义上可解释的样式。与我们之前对"蛙眼告知蛙脑什么"的讨论类似，可以将简单联结主义语义探测器形式化地表述如下： 351

(SCDS)

单元及单元组，能够表征激活它们(或者足够激活它们)的事物。

倘若我们设想，进化机制起到一种"程序员指派"的作用，那么就可以认为，感觉和运动节点的表征位势(representational potential)(语义)受基因决定，它们之间的内在连接强度可以看作是反映了基因序列对于重复出现的环境构造的反应。根据这种观点，这一系统从环境中获取某种数据的规律并对其进行编码，然后对这些规律进行分类，再运用这些分类或多或少成功地作用于外部环境。对于新增的经验，系统会修正连接强度以适应新的环境，并且修正后的连接能够更加准确地作用于先前已熟悉的环境。克拉克(Clark,1993:190)认为，联结主义机器需要具有"基础操作模式，涉及：数据驱动学习、分布式表征的形成、迭加存储技术以及依据语境进行表征和检索"。斯莫琳斯基的 PTC 具有所有这些特征。

12.6 结构分类(Ⅱ)

在前面，我们以**存储**和**控制**为维度(回顾第 6 章关于这些维度的解释)划分了计算的结构。我们已经看到，计算模式也可以基于表征进行参照。如

果一个表征,(1)不止一个元素(触发器、节点等)对它进行编码,以及(2)对这个表征进行编码的元素,也能够对其他表征进行编码,那么这个表征是**(完全)分布式**的。如果一个表征,它只能满足上面的条件(1),也就是说,不止一个元素对它进行编码,但是不满足条件(2),那么这个表征是**准分步式的**。如果一个表征,只有一个元素对其进行编码,那么它是局部定位的(local)。如果机器的转换原则由获得语义解释的状态而确定,那么表征是**语义有效的**,否则是**非语义有效的**。正如斯莫琳斯基所说:"或许对两种程式进行必要的比较,能够得出附属于形式模型的语义解释。在符号的研究取向中,符号(最小的)被用来标识语义上可解释的实体(概念);在定义系统的规则内,那些相同的符号是符号操作程序控制的对象。语义上可解释的实体**同样是**按照定义系统的形式规则所控制的实体。在次符号程式中却不是这样。语义上可解释的实体是系统中大量单元的**激活样式**,而由形式规则操作的实体是在工作网络中单元的个体激活。这些规则运用的是激活传递规则的形式,其特征与符号一操作规则有本质不同"(1989:54)。正如我们已经指出的,在一个(分布式)联结主义系统中,概念层的解释由激活了的单元矢量或样式层指派,而转换原则由节点和连接层(激活一传递原则,连接权值)的状态确定。但在传统数字计算机中,概念解释层与转换到下一种状态的转换层是相同的——都是程序层。以 SHRDLU 为例,放入盒中一个金字塔,是因为执行了这样的命令:PUT(IN(BOX, PYRAMID))。从其中,我们看到了语义解释的对象(思考 BOX, PYRAMID, PUT),使机器执行了这个行为,之后进入另一种程序状态。增加表征维度使我们拓宽了对结构的分类,使之能够包含联结主义结构(见图 12.6)。

352

	控制	存储	表征
图灵机	局部定位	间接	• 准分布式 • 语义有效
冯·诺依曼机	局部定位	• 直接 • 形式寻址	• 准分布式 • 语义有效
产生式系统	分布式	• 直接 • 内容寻址	• 准分布式 • 语义有效
鬼蜮模型	分布式	• 直接 • 内容寻址	• 局部定信(?) • 语义有效
联结主义模型	分布式	• 直接 • 内容寻址	• 分布式 • 非语义有效

图 12.6 结构类别

图中只是大致地对不同机器模型的特征进行了说明，但似乎也能够看 353
出从图灵机到联结主义机器都存在着控制、存储和表征的框架。当我们从左向右看这个图时，机器的总体属性就从局部定位元素的交互作用（更多的相互合作，较少的单独控制）中呈现出来，其他的重要属性，如容错性（fault tolerance）和功能衰减（graceful degradation）都可以看作承载在低层次特征上——尤其是分布式控制和分布式表征。如同丹尼特（Dennett）在与我们稍有不同的语境中所说的："请注意，从冯·诺依曼机的结构到……产生式系统和（带有精细颗粒层的）联结主义机器的结构，在这一进程中发生了什么变化。这些变化或许可称之为平衡力的转移。在由具有较少不连续点的数据描绘的轨迹上运行的、固定的、事先设计好的程序，都被灵活的——非常易变的——系统所取代，这个系统的后续行为，更多是由系统目前正在发生的经历和先前已发生的经历之间的复杂相互作用而产生的结果"（1991：269）。我们将在下一节讨论这些特征。

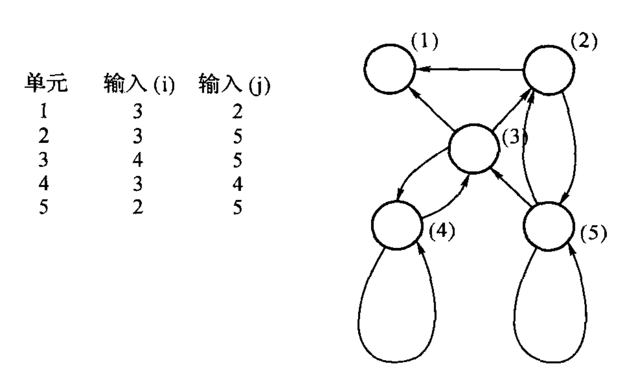
附录　联结主义与图灵的非组织机器

在 1969 年出版的图灵的遗著（成稿约在 1948 年）中，他提出了**非组织机器**（unorganized machines）——"其结构具有很大的随机性"——直到现在，这种机器才开始广为人知（见 Copeland and Proudfoot，1996）。图灵提出了这种机器的三种类型，他分别称其为"A-类型"、"B-类型"、"P-类型"。我们在这里并不关注这三种类型机器的技术细节和它们之间的差别。我们感兴趣的是，它们的一般特征以及图灵所设想的它们对于未来智能研究的重要性。

A-类型非组织机器

这种类型机器的特征，如图 12.7 所示。在这个简单的机器中具有 5 个单元，每个单元都有两个单元连接输入，如左边表和右侧图所示。每个单元要么开，要么关。激活－传递的规则是：两个输入项的乘积减去 1，即：新值＝（输入 1×输入 2）－1。存在一个中央时钟（central clock），能够使所有的激活－传递同步。赋值给每个单元 0 或 1 的结果，如图 12.7 底部所示。例如，在第一个条件下，单元 1＝1，因为它的输入单元（＃2，＃3）的值为：＃2＝1，＃3＝0。运用激活－传递规则我们得到：1×0＝0，1－0＝1。所以单元 354

1 可以获得值 1。对于 A-类型机器与脑的关系，图灵评论道："A-类型的非组织机器，非常有潜力成为包含随机排列神经元的神经系统的最简单模型"(1948/69:8;Ince，1992:Ince，120)。在这里，我们可以看到，图灵已经明确地预见到了模拟真实神经结构的可能。



单元	输入 (i)	输入 (j)
1	3	2
2	3	5
3	4	5
4	3	4
5	2	5

整个机器的 5 种可能连续条件序列:

1	1	1	0	0	1	0
2	1	1	1	0	1	0
3	0	1	1	1	1	1
4	0	1	0	1	0	1
5	1	0	1	0	1	0

图 12.7　A-类型非组织机器(来源:Turing，1948/69：10；Ince，1992：114)

B-类型非组织机器

在图 12.7 中，较简单的 A-类型机器的每一个连接或者说标有箭矢的线段，都被一个环路取代，这个环路如图 12.8 所示。所有 A-类型机器的连接被 B-类型连接取代之后，机器的结构就是图灵所说的 B-类型机器(见图 12.9)。

355

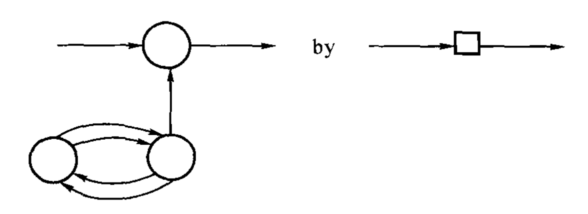


图 12.8　B-类型环路(来源:Turing，1948/69：11；Ince，1992：115)

对于 B-类型机器与脑的关系，图灵评论道："脑只是具有 B-类型机器普遍特征的一种特例……换言之，若这种机器具有适当的初始条件，它们将会

完成任何需要的工作，只要给它充足的时间以及提供足够数量的单元。尤其是，具有足够单元的B-类型非组织机器，它可以找到使其成为带有特定储存能力的普适机器的初始条件”(1948/69:9;Ince, 1992:119)。这里，我们会有一些想法，脑可能就开始于随机的一组连接，那么通过对B-类型机器进行合适的训练，它就有可能进化为一种普适图灵机，因此我们可以将B-类型机器看作是认知结构的一种可能的认知结构。遗憾的是，图灵并没有勾画出他所预见的这样的发展应该如何展开。

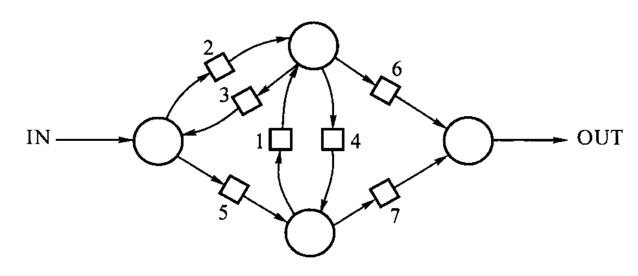


图 12.9 B-类型机器图解(来源:Turing, 1948/69: 15; Ince, 1992: 119)

P-类型机器

P-类型机器是图灵学习实验的基础。图灵明确地认为，人类的学习部分地是以快乐和惩罚(因此称为“P-类型系统”)[4]为基础的:“人类孩童学习，在很大程度上依赖于他自身带有的奖赏和惩罚系统。这说明，有可能只通过两个调节输入就能来执行这个组织，一个调节输入是‘快乐’……而另一个是‘痛苦’或是‘惩罚’。人们有可能会设计出大量的‘快乐一痛苦’系统”(1948/69:17;Ince, 1992:121)。

“P-类型”系统的一般特征如下:“P-类型机器可以认为是没有磁带的LCM(逻辑计算机器，如图灵机)，而且对它的说明在很大程度上也是不完整的。当它具有了相对完整的结构时，其行为并非是确定的，而是会对缺少的数据作随机选择，一旦发现有合适的填补数据，就会对填补数据进行说明、试验和运用。当痛苦的刺激出现时，所有试验的填补数据就会被取消;而当快乐的刺激出现时，它们就会保持恒定”(1948/69:10;Ince, 1992:122)。所以，机器的构造允许机器当它痛苦时删除正试验着的部分结构，快乐时则对其保持(强化?)。图灵也预见到了近来人们所说的联结主义机器可能在数字机器上实现模拟:“一些电子机器在它们实际运行时，我希望它们能够做到这一点[测试多种学习程序的‘教育方法’]。很容易创造一种适合于任何 356

特殊机器的模型，这种模型可以在 UPCM[图灵本人使用的缩写，代指用于实践的通用计算机器(universal practical computing machine)——今天的数字计算机]中运行，而不是如现在一样只能运行在纸面上的机器。如果人们同时确定了可以适用于机器的'教学策略'，那么也可以在机器中对这些教学策略进行编程。人们如果使整个系统能够对自身进行感受，那么，机器对自身的感受会逐渐成为'学业督导员'，能够检测机器取得了怎样的进步。同样地，类似于 A-类型和 B-类型的非组织机器也可以通过它们对自身的感受确定我们对它们所取得的进步。所有这些涉及的工作已经远远超出了单纯纸面一机器的范围"(1848/69:11;Ince, 1992:125)。我们再一次赞叹，图灵远远地超越了他所处的那个时代，我们必须承认他同时是图灵机的发明者，现代寄存机器发展的贡献者，以及联结主义机器及其能够进行学习的预见者。

注释

[1] 萨伽德(Thagard,1986)以及斯莫琳斯基(Smolensky,1988a)的论题(1i)提出，联结主义可能为心一身难题提供一种新的视角。但因为联结主义机器是从物质上抽象出来的(它们可以在神经元和硅上实现)，也就很难确定联结主义自身对于心一身难题究竟有何帮助。

[2] 斯莫琳斯基经常称之为有意识的规则"解释"，虽然这在计算机科学范围内是标准的术语，但是对于非计算机科学领域的人来说，这种说法似乎给人的印象不是对规则进行"解释"，而是"解释"本身的规则，像是在说关于"解释"的法则。

[3] 这里，"单调"意味着**递增**。传统的演绎推理是单调的，因为从给定的前提中推演出的一组结论总是增加的，不会缩减。但是在非单调推理中，结论能够缩减，能够给予附加论据。

[4] 我们不能忘记 1948 年是行为主义的全盛时期，而他 1950 年的论文提出"图灵测试"，就具有显著的行为主义式的学习特征。

【思考题】

CCTM 的理据

从人类的表现行为中能够得出哪些理据？

从人脑的结构中能够得出哪些理据？

联结主义与联想主义的历史关联

357 联想主义工作网络与联结主义工作网络有哪些相似之处及不同之处？

联想主义是联结主义的一种特例吗？

联结主义的解释:PTC

斯莫琳斯基对于心一脑的描述而区分出的三个层级是什么?

在联结主义模型中我们如何表征"盛有咖啡的杯子"?

联结主义表征所具有的两种构成类型是什么?

斯莫琳斯基区分的两种认知能力("知识")是什么?

PTC 有哪些构成要素?

PTC 认为次概念层与神经层的关系如何?

PTC 认为次概念层与概念层的关系如何(对于有意识的规则应用和直觉加工)?

系统具备认知的必要条件是什么?

活动样式具备怎样的必要条件才能成为一种意识状态?

联结主义网络通过哪两种方法可以获得其语义特征——它的表征能力?

可将联结主义工作网络的语义探测器看作什么?

结构分类(II)

结构分类有哪几个主要维度?

语义有效(SE)与非语义有效(SI)之间的区别是什么?

对于 SE 和 SI 的区分,数字机器与联结主义机器有何不同?

在何种意义上,存在计算机从图灵机器到联结主义机器的谱系? 358

【推荐读物】

CCTM 的理据

关于联结主义的经典论述,参见 Rumelhart and McClelland (1986a)第 1 卷第 I 部分。

联结主义与联想主义的历史关联

关于联结主义与早期心理学发展关系的研究讨论,参见 Valentine (1989)和 Walker (1990)。在 Bechtel (1985), Rumelhart and McClelland (1986a),以及 Ramsey (1992)中,可以找到有关联想主义与联结主义之间关系的零散讨论。

对联结主义的解释

Smolensky (1988a)给出了 PTC 最重要的论述,他的一些主要观点可见 Smolensky (1989)。对 Smolensky (1988a)中关于 PTC 的原始论述的进一步讨论以及斯莫琳斯基的回应,可见 Rosenberg (1990a, b), Mills (1990),以及 van Gelder (1992)。Clark (1993)对联结主义模型的优势和缺点给出了长篇评论。

意识、认知与内容

对于突现问题的考察,参见 Beckermann et al. (1992)。有关进一步将认知

功能看作是有机体适应环境的结果的讨论，参见 Copeland (1993b)第 3 章，尤其是第 3.6 节。关于 SCDS 的更多讨论，参见 Ramsey (1992)。

结构分类

van Gelder (1997)的第 5 节，作出了与本书讨论非常不同的、更具一般性的分类。

图灵的非组织机器

关于这部分内容，还可参见 Copeland (1998)，以及 Copeland and Proudfoot (1999)。

13 心智联结计算理论的评论

13.1 引 言

> 有时，看似肯定是因为一场大火而引起的滚滚浓烟，却只不过是一列游行队伍践踏出来的飞扬尘土。
>
> ——丹尼特(Dennet, 1991:257)

在这一章里，我们将再次讨论 CCTM 最初遇到的一些难题，以及联结主义者们为之提供的回答。这些回答应该被看成是暂定的，因为联结主义毕竟还是一种很年轻的理论，具有非常大的发展潜力。 359

13.2 CCTM 与人脑的差异

显而易见，早期的联结主义理论认为 CCTM 模型与真实的人脑只具有一些间接的关系(见第 12 章，对斯莫琳斯基的 PTC 的讨论)。但我们有所侧重和选择地指出，两者间的一些深层次的不同，也许对于引导我们如何从 CCTM 模型推断出人脑的特征是非常有帮助的。

神经元与单元

例如，克里克和浅沼智行(Crick and Asanuma, 1986: 367-71)就阐述了

联结主义单元与脑神经元的一些明显相似的地方："它们都具有多重输入、某种类别的求和规则、某种阀域规则以及分布在其他多个单元中的单一输出。"但他们又提醒道："如果真实神经元的特征为尝试模拟神经系统工作的人们提供了有用的参考，那么他们就不该将两者的特征混为一谈，事实上，单个神经元与单个单元并不是完全的对应。"最常见的一种解决方案是几组（真实的）神经元对应一个单元，但他们指出："如果能详细地说明或多或少的一些真实神经元，如何能形成某一神经元组，那么神经学家或许才会接受这一点。但这样的解释，即使有也是少得可怜。"他们继续指出，人脑并不总是具有联结主义模型的表现，如果联结主义模型真实地对应着人脑神经元的结构，那么，可以开列一张联结主义模型单元应当具有的设置清单，"那些设置，[CCTM]理论家们之所以钟爱，是因为他们只从文字上进行解释，而这些设置事实上并没有任何可靠的实证基础"：

360

1. 一些单元能够对某些单元产生兴奋，而对另一些单元进行抑制；
2. 单元仅接受某个特定单元的兴奋，且其输出也仅对某一特定单元产生抑制；
3. 单元与同类型的所有其他单元相连接；
4. 仅靠单元本身，就能激活其他单元。

我们已经知道（见第 9 章）人脑中存在着极其繁多的神经元种类，而一个特定的 CCTM 模型通常只具有一种类型。因此，我们可以再加上一条：

5. CCTM 模型通常只包含某一类单元，而人脑则包含很多种类的神经元。

化学物质

1. 在第 3 章末，我们已经注意到大脑使用神经递质和神经调质，"改变细胞的功能，使神经网络急剧地转换它的整个活动模式"（Arbib，1995：6）。CCTM 模型没有使用任何相类似的事物。

几何构造

1. 正如我们前面（第 10 章、第 11 章）讨论的，许多联结主义的模型都有很多不同的功能层次（数量可随意增加）。但是，大脑似乎有着更为复杂的物理几何分层、连接和投射，它们的计算作用可能还无法复制到现有的模型

中。例如，脑皮层的垂直柱状结构，每一层内都有着极其繁多的连接。

2. 我们也不应该忘记(见第 3 章、第 8 章)，不同的脑区似乎都至少部分 361 地参与了某类计算，如布洛卡区和维尔尼克区。

学 习

1. 脑的学习似乎不需要过多的重复或监控。

2. 脑似乎没有反向传播的生理机制，而在 CCTM 模型中反向传播由主计算机(host computer)执行。

规模比例

它们之间当然还会存在规模上的差异，这个问题尽管在理论上似乎不成问题，但它们之间相差的程度实在太令人震惊了。比如，丘奇兰德估计大脑中约有 10^{11} 个非感觉神经元，且每个神经元平均约有 10^{3} 种突触连接(Churchland，1989，第 9 章)。

> *激活矢量*：假设大脑有一千个子系统，每个子系统工作时有 10^{8} 的容量(包含 10^{8} 个单元)——一个包含 10^{8} 个单元的矢量够写出一整本书。从 10^{8} 个单元中可以建构多少种不同的矢量呢？如果每个单元都带有(保守估计)10 种不同的值，那么工作时将有 $10^{100,000,000}$ 种不同的激活矢量。这究竟是个怎样的数据呢，史蒂芬·霍金(Hawking，1988)估算说"在我们现在能查明的宇宙范围内，大概有[10^{80}]个粒子"，与 $10^{100,000,000}$ 相对比，10^{80} 实在太小了。而且这只是我们假设的一千个子系统中的一个子系统的激活矢量的数字。

> *权值与连接*：如果每一神经元平均有 10^{3} 种连接，那么每一子系统就有 10^{11} 种连接，而每一种连接都有 $10^{10,000,000,000}$ 种矢量解释。

这意味着，人脑蕴藏着无比巨大的对所表征内容能够作出精细区分的能力。有一点值得我们注意，尽管我们可以设计一种工作网络，用它来模拟人脑的工作，但如果这样一种设计只是我们试着正确认识认知的副产品，那么，这将意味着，脑的结构并不是偶然地与认知结构相关，那会变得更加 362 有趣。

13.3 CCTM:联结主义的优点

福多和派利夏恩(Fodor and Pylyshyn, 1988)(下文简称 FP)列举了十多个被普遍接受的理由,这些理由使一些人相对于他们称之为的“经典”结构或“传统模型”(DCTM)更偏爱于联结主义(CCTM)结构。其中大多数理由我们前面已有所涉及(1988:51-4):

联结主义具有的 11 种优点

1. 认知过程的速度与神经元速度相关:“一百步”限制。

2. 传统结构很难具有大容量模式识别和基于内容的检索能力。

3. 传统计算机模型不能统合对“规则控制”行为和“例外”行为的解释。

4. DCTM 在处理非语言或直觉过程上,无法获得进展。

5. 传统结构对受到的损坏和干扰,极度敏感。

6. 传统结构的存储是被动的。

7. 传统系统的规则基础将认知描述成“有一或一无”。

8. CCTM 模型对于不同规则的适用程度连续变化。

9. CCTM 对人类行动的非确定性作了更好的模拟。

10. 传统模型无法出现功能递阶衰减的特征。

11. 传统模型受现代计算机的技术特征的影响,对神经科学的成果很少或完全没有涉及。

福多和派利夏恩想要说明,以上所列举的赞成联结主义的理由其实是无效的,因为这些理由都有下面这些那样或这样的缺陷:

福多之叉(Fodor's fork)

1. 上述批评理由所指向的,**并非是经典认知结构的本质属性**,或者

2. 所指向的是**实现层**或神经层,而不是认知层。

五类优点

363 FP 将最常见的支持联结主义结构的理由总结为五类,然后逐一进行了回应。

1. 并行计算和速度问题(实现层)

这一部分主要针对两个目标。第一个反对的目标是费尔德曼和巴拉德(Feldman and Ballard，1982)的“100 步”规则(“100-step” rule)。费尔德曼(Feldman，1989:1)这样阐述这条规则:“人脑是一种与传统计算机截然不同的信息处理系统。人脑的基础计算元件在毫秒范围内运作，这大概要比现在的电子元件慢 100 万倍。可是，人脑对于某种复杂任务的反应时间仅需几百毫秒，因此系统要处理很难的识别问题，必须限定在大概 100 个计算步骤内。但由于这个时间限制，一个神经元只可能给另一个神经元传递一个简单的信号。”FP 用三段论的形式解读“100 步”规则:

(P1) 神经元的激活需要几毫秒。

(P2) 相关的认知任务在几百毫秒内发生。

(C) 因此，对这些任务的算法分析只能在百步内完成。

FP 对这个论证的回应是:“从联结主义典型的讨论方式看，这个问题与传统的认知结构的充分性是无关的。比如，‘100 步限制’明显指向的是实现层。这一规则只能排除这个(荒谬的)假设，即认为认知结构在人脑中的实现方式与它们在电子计算机上实现的方式完全一样”(1988:54-5)。但是，CCTM 的支持者可能会说，如果硬件是产生因果关系的机制，而且被限制在几个毫秒内运行，那么两种心理状态对应着的两种物理状态之间的转换速度，**不能比计算状态转换的速度快**。因为用物理属性描述的神经系统是在几十微秒内传递信息，所以有因果关系的连续神经状态间的转换不能超过这个速度。因而，计算机内计算状态间的转换也不能超过这个速度，所以计算过程必须被限制在 100 个连续的步骤以内。因此，联结主义支持者认为，即使“100 步限制”只是“实现问题”，也不能改变上述事实。

第二个反对联结主义的理由，涉及这个结论，“并行计算机网络的论证本身既不能用来反对传统结构，也不能支持联结主义结构”(1988:56)。这是因为“尽管在 VAX[①] 上运行的大部分算法是序列的，但在实现层，这样的计算机还是存在着‘极大程度的并行’加工过程;事实上，整个电子装置几乎处处都同时进行着电子活动”(1988:55)。“传统结构绝没有在任何意义上，排除并行执行多重符号处理……见……Hillis(1985)”(1988:55-6)。但是 CCTM 支持者可以作出两个回应。首先，VAX“极大程度的并行”与联结主义的并行所说的并不是同一件事。分布于 VAX 内的电子活动，与联结主义

364

① VAX(virtual address extension)，即虚拟地址扩展的计算机体系。——译者注

网络的激活扩散不同，因为分布的电子活动并不能直接获得语义解释而确定对象。其次，希利斯(Hillis)的连接机器[①]要远比联结主义机器复杂得多，其激活传递值分布在 0,1 之间。

2. 对干扰和物理损坏的阻抗(实现层)

FP 开头这样说道："单元的分布式联结，只有能够满足其表征也是神经元分布的，才能具有损坏阻抗(damage-resistance)的功能。然而，表征的神经元分布与经典结构的相容程度，与其与联结主义网络的相容程度是一样的。在经典结构中，将内容分布存储于物理空间内，所需的仅仅是存储寄存器。"(1988:52)联结主义与经典模型在损坏阻抗上，存在两个主要区别：首先，前面提到分布式表征有两个关键特征，而传统模型只具有第一个特征：

(DIST)
(i)R 是分布式表征，当且仅当
R 实现于多个单元；
(ii)参与实现 R 的单元，同时也能实现大量其他表征。

其次，经典结构通过自动复制(redundancy)而获得损坏阻抗，即在整个机器内存储多个复制的表征。但联结主义结构并不需要通过将多个复制的表征分散于整个网络而获得干扰和损坏阻抗。两者的不同之处体现在，分散在经典模型中的多个复制的表征，如果其中的一个受到了损坏就将其剔除，而不会影响系统存储的其他复制的表征。但在联结主义模型中，由于具有上述特征(DIST ii)，如果储存信息某个片段的单元受到损坏，那么所有其他存储有关这个信息片段的单元也会受到损坏。

365

3. "软"约束，连续量值，或然机制与激活符号(非本质属性?)

FP 这样说道，"在经典的规则系统里，确定哪个规则需要激活，取决于经典模型的功能结构，且依赖于连续发生变化的量值(varying magnitudes)。事实上，这正是已实现了的'专家系统'所做的，如在产生式系统的规则解释器中应用了贝叶斯原理。基于规则加工的'软'或者或然(stochastic)特征，或者源于确定性规则与具体执行时的实际数值的相互作用，或者源于与噪音输入或信息干扰传递的相互作用"(1988:54)。这样的系统与联结主义系统的一个明显区别是，这些系统能够读取和遵循贝叶斯概率公式。但是读取和遵循贝叶斯概率公式，并不会使系统像使用连续发生变化的激活值和

① 连接机器(connection machine)是一种超并行计算机。——译者注

权值的联结主义系统那样,“依赖于连续发生变化的量值”。FP 还指出当下经典心理模型所面临的一个难题是,“功能衰减可能是由于它们整体上缺乏智能而导致的特例:在有限的方法都失效后,它们只是不具有足够的智慧知道接下来该做什么”(1988:54)。但是,这似乎并不是联结主义网络处理缺省输入的**方式**;使之功能衰减的结构特征,同样也必然能使系统获得自动修补并使之具有自动归纳功能。

4. 规则的明晰性(非本质属性)

FP 评论道:“人们还**可以**通过指出联结主义结构不具备认知加工规则的明晰性而反对它,因为联结主义结构的定义,**排除**了各种逻辑一句法功能,而在经典结构中,这些逻辑一句法功能是用来对规则以及应用于规则的各种执行机制进行编码的”(1988:57)。我们还是不清楚,为什么联结主义模型的“定义”与规则的明晰性不相容。斯莫琳斯基(Smolensky, 1988, 第 6 部分)就提出了一种方法,使联结主义网络的规则明晰性并不比产生式系统差。

5. 关于“脑一式样”模型(实现层)

FP 的下一个观点是:“我们有理由怀疑上述所列的各种属性(关于神经元和神经活动的生物事实),是否或多或少直接地反映在执行逻辑推理的系统结构中……这里强调的是,系统‘高层’结构与系统‘低层’结构不具有同构性,甚至也谈不上相似。对人脑结构的推断,总是以一种过于直接的方式被认知结构假设所采纳”(1988:58-60)。我们还不清楚 FP 对于“脑一式样”模型的这些评论具有怎样的效力。但“脑一式样”模型的关键点就是,**在其他条件相等的情况下**,我们应该选择更符合于神经运行的那个理论,而不是别的。 366

FP 在论文结尾处写道:“许多支持联结主义的观点最好应该这样措词:认知结构在某种(抽象‘单元’的)工作网络中**实现**。只有以这种方式理解他们的那些观点,才能在认知结构是什么的问题上保持中立”(1988:60-1)。

对上述 CCTM 五类优点及回应的一般性评论

可以看到,传统结构通过使模型**复杂化**,而展现出上述 1—5 特征——它们并非从其内在结构中自然地突现出来。而且,FP 似乎并没有在回应中全

部涉及 CCTM 起初的 11 种优点(尤其是 2、4、6),而这些优点都与联结主义结构的并行分布特征相关。最后,CCTM 似乎还有一些优点并未列出,比如,它能够进行自动归纳、原型提取、修复损伤,还具有能够进行迅捷再学习以及类似于人的学习等优点。麦克劳克林和沃菲尔德(McLaughlin and Warfield, 1994)对 CCTM 在学习方面是否真的具有优势进行了讨论。他们考察了一些关于 DCTM 的学习类型的研究,并在速度和准确度方面与反向传递做了比较,最后得出结论:"现在还没有足够的证据证明,联结主义结构比传统结构在模拟模式识别功能和通过学习习得这些功能上,要更加出色"(1994:392)。

对实现层的一般性评论

FP 用"福多之叉"的第二个分叉反对 CCTM 的 1、3、5、7—11 优点(我们将这些称为"I 类优点"),他们的这种策略很可能会适得其反。原因是,根据 FP 的观点,人们对认知结构的分歧就等同于对表征层的加工状态的分歧,
367 而表征层的加工状态是指有机体(对外界状态进行编码后)表征层的状态。心理解释针对的是表征状态。我们直觉上认为"I 类优点"似乎非常符合人的**心理**特征,而 FP 却要从**实现**层而不是从认知结构层对其进行解释,即他们认为,认知结构层并不属于认知理论的一部分。因此,我们可以对福多和派利夏恩的"福多之叉"进行改写:

经典结构的两难选择

这些特征,要么被证明不是真实的认知特征(需要论证),要么它们是认知特征,但无法得到认知(计算的和表征的)解释。

经典结构两者只能选其一,但无论选择哪一个,都不能令人感到满意。而联结主义却并不需要做这一选择,因为联结主义承认这些现象是认知的,并且能够获得认知解释。最后,我们也不能忘记,联结主义结构是"唯一的具有实现层"的理论,这是联结主义所做出的重要贡献,而这一点恰恰又是传统理论所亟需的。传统理论是一种没有对实现层进行解释的认知理论,没有提出认知如何确切地实现于脑——人脑的观点。就像一种关于生命如何产生和发展的生物学理论,却没有解释有机体如何就具有了生命。实现理论(implementation theory)对于传统理论并非是可有可无的奢侈品——没有实现理论的传统理论,只能是一种计算机或虚假的理论,而不是关于人类及其相关机能的理论。或者我们换个角度看这个问题:如果有人认为传统理论就像是化学,联结理论是传统理论的补充,那么,正如麦克劳克林巧

妙回应的:“如果联结理论是传统理论的补充,那么联结理论就应该是量子力学,而传统理论只能是化学。如果有诺贝尔心理学奖,无疑应该授予对脑的联结网络如何实现了传统认知结构给出解释的那个人”(McLaughlin,1993:184)。

13.4 CCTM与中文体操馆

与中文屋类似,塞尔(Searle,1991)又提出了中文体操馆论证,反对心智的标准数字计算理论与联结主义模型有所不同,但与中文屋论证的效力是不一样的:“并行‘类脑’加工的特征与纯粹的加工计算方面毫无关联”(1990:28)。事实上,人们对中文体操馆存在着两种不同的看法,一种认为中文体操馆是中文屋论证的延伸;另一种认为,中文体操馆针对的是并行与串行机器的计算力。 368

中文体操馆论证

塞尔(1980)提出中文屋论证用于表明,计算既不是认知、心智、意向性等的构成部分,也非它们的充分条件。因而,如果类脑特征的联结主义模型与其计算属性无关,那么联结主义程序(以及数字程序)既非认知、心智、意向性等的构成部分,也不是它们的充分条件。塞尔这样说道:“想象有一座中文体操馆:馆内有许多只会说英语的人,这些人的活动与联结主义结构中的节点和突触一样……馆内没有一个人会讲中文,哪怕一个中文词汇,那么包含了所有人的整个系统也就无法学会任何中文词汇的意思。但是,经过适当的调整,系统可以针对中文问题给出正确的答案”(1990b: 28)。“无论系统以串行还是并行的方式运行,仅靠形式的计算,系统都无法获得包含丰富思想的语义内容;这也是为什么中文屋论证能驳倒任何形式的所谓强人工智能”(1990b: 28)。塞尔的中文体操馆论证,事实上并未像他的中文屋论证那样具有相对较强的模拟力。人是节点,但什么是激活层次、激活一传递规则、联结、联结强度呢?什么是矢量间的相乘运算?输入又如何转换为输出?究竟体操馆的什么对应着联结主义模型的这些特征,并不清楚,而这些对应关系又非常重要,所以这一类比有些牵强,不能与中文屋和数字机器的类比相提并论。

但是,假设我们已经有了较为合理的联结主义机器模型,那么这个论证

会是怎么样的呢？可能是这样的：

(1)中文体操馆模拟了联结主义机器。

(2)中文体操馆不会说，也不懂中文。

(3)所以，联结主义机器不会说，也不懂中文。

这一论证似乎具有这样的形式：

(1)M模仿了P。

(2)M不具有性质F。

(3)所以，P不具有性质F。

但是，正如塞尔自己所坚持的，这种推理形式是无效的——看这个例子：

369 (1)M模仿了P(P是一种哺乳动物，它正处于哺乳期)。

(2)M不具有性质F(不产奶)。

(3)*所以，P不具有性质F(不产奶)。

因此，塞尔不能因为中文体操馆不会说/不懂中文而得出结论，认为联结主义模型不会说/不懂中文。

"串行－并行"论证

塞尔为得出结论"并行加工不能摆脱中文屋论证"，似乎考虑(论证?)到了以下两点(1990b:28)：

(A)*联结主义程序在串行机器中运行*："…因为并行机器还很少，所以联结主义程序通常在传统串行机器中运行。因而，并行加工也不能摆脱中文屋论证"(1990b：28)。

(B)*串行与并行机器(弱)等效*："从计算角度看，串行和并行系统是等效的：任何能在并行系统内完成的计算，同样也能在串行系统中完成。中文屋内的人从计算角度看都与这两者等同，又因为仅靠计算无法懂得中文，那么无论是串行还是并行系统都无法懂得中文"(1990b：28)。

以上两个论证似乎具有这样的形式：

(A)联结主义程序通常在传统串行机器中运行。

(C)因而，中文屋论证也同样适用于并行联结主义机器。

(B)任何能在并行机器中计算的函数，也能在串行机器中计算。

(C)因而，中文屋论证也同样适用于并行联结主义机器。

即使(A)和(B)在没有任何限定性条件下是正确的——那么,就能得出结论(C)吗?不能。强人工智能只是宣称"适当编程了的计算机"具有认知,(A)和(B)最多也只能表明串行和并行机器弱等效。运行不同程序 370 的机器能够计算相同的函数,即便假设在没有任何限制性条件下(A)和(B)正确,论证也得不出结论(C),况且(A)和(B)如果正确,也不能没有任何限制性条件。

对于(A) 联结主义模型中至少具有三个层次[1]:

> 层次1:离散的,连续串行近似的,并行模型(能在串行数字计算机中运行——如Jets & Sharks网络)。

这样并行模型近似于:

> 层次2:连续的,连续并行模型的,并行现象。

这些现象准确地模拟:

> 层次3:连续的,并行现象自身。

某些联结主义模型是离散的,能够通过在串行数字机器中运行进行模拟,但是绝大多数联结主义模型是连续的,能在串行数字机器中运行是因为离散近似(discrete approximation)(再如Jets & Sharks网络,第10章)。联结主义假设最初的连续模型,其重要心理特征(如工作网络如何进行学习以及学习什么)可通过数字近似(digital approximation)而保存。最重要的是,在串行数字机器模拟中的联结主义模型的心理相关特征,是串行数字机器联结主义模型虚拟层的特征,而不是串行数字机器的基础特征。我们不能从一个层次的特征属性推导出另一层次也具有同样的特征属性——就模型自身而言,它没有更低的层次了。

对于(B) 塞尔似乎确信,以并行方式集合在一起的几台图灵机,与单个(串行的)图灵机在计算上等效。人们业已证明,任何能被多个并行运行的图灵机执行的计算,也能在单个图灵机中执行。但联结主义机器并不是多台并行运行的图灵机,因而(B)什么也说明不了。

结 论

塞尔似乎把联结主义看成是另一种类型的强人工智能,另一种无论如何都无法完成的程序"实现"理论。但有两点需要注意:首先,许多联结主义 371

者(见 Smolenksy，1988a；Rumelhart，1989)并不认为联结主义是标准程序的实现理论。因此，即便经典理论认为认知结构对于认知至关重要，联结主义的认知结构却与之有所不同。其次，联结主义者中很少有人认为，**通过模拟或对认知建构模型就能够复制认知**。尽管联结主义者，如斯莫琳斯基(Smolensky，1988a)，宣称他们的模型要比神经元更抽象，因而这些模型或许可以用其他物质材料实现，而不同意塞尔的观点，即认为只有某些物质材料才可能复制认知系统。这是经典理论与联结主义理论的一个重要区别：联结主义模拟的认知活动是对神经活动细节的抽象，并且就能自然而然地坚持认为，真实的人脑(或等效物)对于认知活动是必要的。而经典理论者若采取这样的立场就显得很不自然，因为他们的理论与真实神经元结构，如同与微波炉结构的关系一样，毫无关联。因此，说联结主义模型与神经结构的这种等效，不等于说联结主义信奉强人工智能，可能会遭到反对，如塞尔。但塞尔的反对论证与这一点并无必然关系，因为除塞尔的观点和强人工智能之外，还有第三种选择，即物质＋程序＝认知，或更形象地说，物质提供了认知的“材料”，而程序提供了“形式”——从而能够形成各种思维结构。

13.5 CCTM 与命题态度

到目前为止，我们对于 CTM(同时包括 DCTM 和 CCTM)的讨论，可获得以下三个论题：

(R)
确实存在着所谓的命题态度(态度的实在论(Realism))。

(C)
态度的核心特征(Central features)可从大众常识心理学中获得。

(N)
态度的(科学)本质(Nature)可通过认知科学研究而得到解释。

在这一看似合理的图景中，先是运用大众常识心理学，将态度看作是认知科学的研究对象，之后认知科学给出关于态度本质的、精细的、科学的解释(想一下我们的常识知觉如何把水分辨为清澈、无味的液体等等属性，之后科学才能在此基础上得出这种清澈无味的液体是 H_2O)。但是将这三个论题整合在同一理论框架中，却并不完全稳定，或没有自相矛盾的地方。如，某人是态度的实在论者(R)，他认为态度可从常识心理学中分辨出来

372

(C),但否认认知科学能够对其进行解释(或许它们并不能是自然科学能够研究的对象,而仅能依靠思辨对其解释):

> **(J)**
> 认知科学的任务(Job)是解释认知,但态度不属于认知科学解释的范畴。

或者,也可能有人是态度的实在论者(R),且相信认知科学能够对其进行解释(N),但认为(C)和(N)可能或必定存在冲突,而最终(N)必定或应当获得胜利:

> **(S)**
> 态度的核心特征必须经由科学(science)进行探索,而不是由大众常识心理学提供给我们。

或者更激进地,有人可能是取消论者,否认(R),否认命题态度的存在——它们就像巫术和生命力一样,只是大众常识理论的错误臆造:

> **(E)**
> 事实上,不存在命题态度(态度的取消论(eliminativism))。

我们在这里提出这些争论的目的,并不是试图给出解决方案,而是为讨论涉及 CCTM 对于态度问题,应该采取实在论还是取消论的争论作铺垫。这种争论是更值得关注的,且直至今日依旧持续。

通常关于每一命题态度的观念,似乎都具有某种具体特征,诸如由多个概念构成,可在语义上评估其真或假(满足条件或不满足条件,等等),是离散的,具有因果效力等等。数字机器似乎特别能体现这些特征,因而命题态度的这些特征的存在,似乎支持了 DCTM 比 CCTM 优越的观点。但也未必,可能更糟。有人主张,因为 CCTM 无法兼容大众常识的命题态度,所以倘若认为联结主义是对的,那么命题态度就不存在——取消主义。如果联结主义是正确的,那么联结主义必须,要么证明命题态度并不真正拥有这些特征,要么证明联结主义模型无论如何都能兼容这些特征。

DCTM、命题态度与常识心理学

在前面的章节中已概述了 DCTM 关于命题态度的观点,即每个命题态度是关于某一表征的某些特殊计算关系。我们形象地总结说,表征储存在适当"盒"内——比如,对于'红酒有益健康'的**相信**,涉及在**相信盒**内储存了 373

一种心理表征，其与‘酒是健康的’发生作用（例如，**红酒有益健康**（RED WINE IS HEALTHY））。我们注意到，根据这种理解，每一计算状态都包含一组重要特征：

1. 它们部分地由**表示范畴的术语**构成（比如，**酒**（WINE））；
2. 它们**在语义上是可评估的**：关于某事物为真或假等等；
3. 它们**在功能上是离散的**，去掉一个表征不会影响其他表征（回顾谓词逻辑、语义网络和框架）；
4. 它们对于产生其他命题态度和最后输出结果，具有**因果作用**。

计算状态的这些特征似乎同样也符合关于态度的常识概念的特征，如，信念：

1. 信念部分地由**概念**构成；

下面的常识概念的特征称为“命题模块”（Ramsey, Stich, and Garon, 1991）：

2. 信念**在语义上是可评估的**：关于某事物为真或假等等；
3. 信念**在功能上是离散的**，去掉一个相信不会影响其他相信；[2]
4. 信念对于产生其他态度和最后的行为结果，具有**因果作用**。

拉姆塞等（Ramsey et al., 1991）提出，态度的因果作用极其多变，并且满足克拉克（Clark, 1993: 194）所提出的“等势情境”。

> *等势情境*：一个人可以拥有两对特定的，**能够引起某一具体行为**（或引起其他的相信—渴望）的相信—渴望，但是在特定的情境中，**事实上**只有其中一对相信—渴望起作用。

374 对此，他们列举了两个例子进行说明，其中的一个例子是，假设我们想要解释为什么爱丽丝在一个特定情境中去了她的办公室。假设她既想查看电子邮件，也想与研究助手交谈，而且她认为（只有）在办公室里才可以做这两件事：“常识心理学认为，爱丽丝去办公室，可能是由其中的一对信念—欲望，或者同时两对而引起。至于究竟最终是哪一种起作用，由当时的特定情境所决定”（Ramsey et al., 1991: 99）。比如，爱丽丝的决定可能是因为，在某一特定情境中她只具有查看邮件的时间，尽管她想与助手交谈，但在时间上却不允许。通过使程序通达一个数据结构而不是另一个而满足等势情境，因此命题态度在 DCTM 中很容易获得模拟。

CCTM、命题态度与常识心理学

从前面的阐述中可以发现，CCTM 的网络表征具有下面一组重要特征：

1. 它们是**分布式的**；
2. 它们**对语境敏感**（回顾第 12 章“咖啡故事”）；
3. 它们的单元和激活传递属于**次符号层**；
4. 它们是认知现象模型的**认知**成分。

现在的问题是，CCTM 是否能够建构命题态度的常识概念？很明显，DCTM 是能够做到的。有一些学者，如戴维斯（Davies，1991）认为，对态度的常识理解部分地参与了表征的概念建构与因果建构，使表征最终能够形成“思维语言”，进而得出结论，认为由于联结主义符号的语境敏感性，因而“网络不会出现句法和因果的加工体系；而常识图式能够产生句法和因果的加工体系；因而联结主义是反常识图式的”（1991：251）。其他一些学者，如拉姆塞等（Ramsey et al.，1991），不仅强调命题层次，而且也认为 CCTM 是无法建构命题态度的常识意义的，我们将重点阐述其后一种观点。为了更好地说明他们的观点，他们首先提出了一个小的工作网络“A”（如图 13.1）[3]。他们将命题在这个工作网络的输入层进行编码，利用反向传递训练网络，使之能够识别输出节点结果的正误，进而进行自我调整，如图 13.2。 375

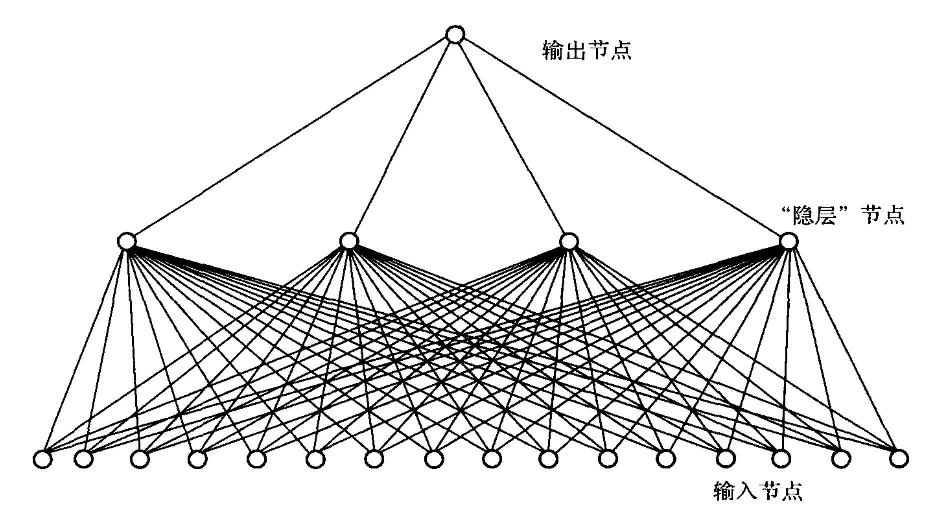


图 13.1　工作网络“A”的结构（来源：Ramsey et al.，1991：107，图 4）

命题	输入	输出
1 Dogs have fur.	11000011 00001111	1 true
2 Dogs have paws.	11000011 00110011	1 true
3 Dogs have fleas.	11000011 00111111	1 true
4 Dogs have legs.	11000011 00111100	1 true
5 Cats have fur.	11001100 00001111	1 true
6 Cats have paws.	11001100 00111111	1 true
7 Cats have fleas.	11001100 00110000	1 true
8 Fish have scales.	11110000 00110000	1 true
9 Fish have fins.	11110000 00001100	1 true
10 Fish have gills.	11110000 00000011	1 true
11 Cats have gills.	11001100 00000011	0 true
12 Fish have legs.	11110000 00111111	0 true
13 Fish have fleas.	11110000 00111111	0 true
14 Dogs have scales.	11000011 00001100	0 true
15 Dogs have fins.	11000011 00001100	0 true
16 Cats have fins.	11001100 00001100	0 true
附加命题		
17 Fish have eggs.	11110000 11001000	1 true

图 13.2　命题和编码表(来源:Ramsey et al., 1991: 107, 图 3)

当网络对命题持续输出的正确率大于 0.9,而错误率少于 0.1 时,训练 376 结束。最后,得出稳定的权值配置,如图 13.3。之后,他们又增加了一个额外的新的命题来训练网络,但发现所有的权值都有细微的变动,有些甚至是发生了戏剧性的变化(见图 13.4)。

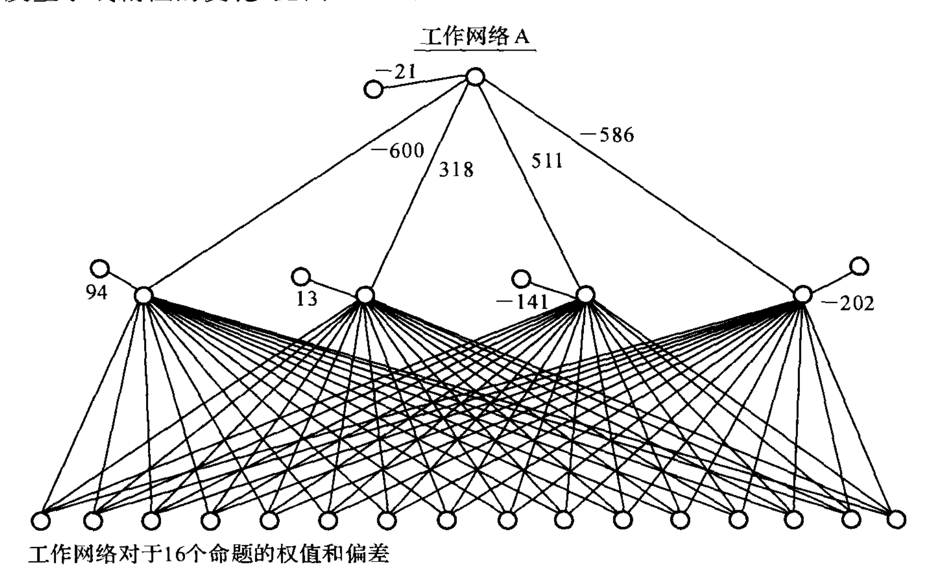


图 13.3　工作网络"A"的权值(来源:Ramsey et al., 1991: 108, 图 6)

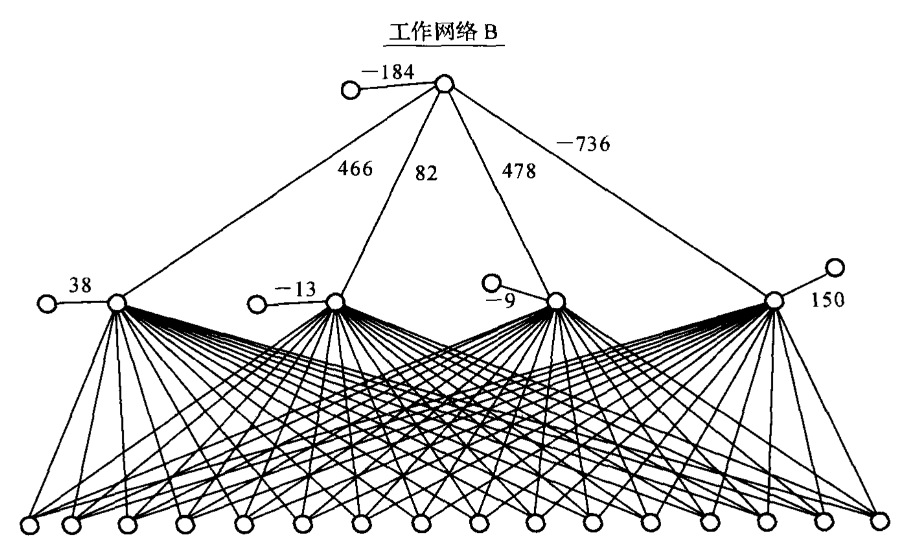


图 13.4　工作网络“B”的权值(来源:Ramsey et al., 1991: 110, 图 9; 图 13.1—13.4 经 Ridgeview Publishing Co. 许可复制)

依据这些发现,拉姆塞等人对联结主义与常识心理学的关系提出了两类论证。

整体性论证(HA: the holism argument)

(1)常识心理态度是命题模块的。

(2)因而,常识心理态度呈离散状态,各自具有不同的因果作用。

(3)而工作网络如“A”,是“整体性”的。例如,它们不具有离散状态,如常识心理态度一样,各自具有不同的因果作用。

(4)因而,工作网络不包含常识心理态度。

这里,拉姆塞等(Ramsey et al., 1991)说道:“联结主义网络……对于任何特殊的具体命题,并没有其独特的呈现状态或者对应具体的网络局部。网络‘A’编码的信息,整体地、分布地储存于网络中……因此,讨论某个具体命题的表征是否在网络的计算中具有因果作用,也就变得毫无意义”(同上: 108-9)。他们继续指出,“常识心理学似乎预设,任何特殊的相信或记忆是否在具体的认知片段中具有因果作用,通常是有明确答案的。但是,如果相信和记忆是由像我们提出的那样的联结主义网络提供,那么这个问题就不会有任何明确的意义了”(同上:109)。 377

自然类论证(NKA: the natural kinds argument)

(1)常识命题态度状态属于心理的自然类。

(2)工作网络编码命题态度的状态不属于自然类。

(3)因而,工作网络不包含常识心理态度。

拉姆塞等(1991)评论说:"显然……存在**无限多**的联结主义网络可以像网络'A'那样,表征'狗有皮毛'这个信息……从联结主义模型的构建者角度看,能够模拟认知主体相信'狗有皮毛'的所有这些网络,都不是真实的那一类,而仅仅是混乱的不连续组集。常识心理学把人相信'狗有皮毛'看作是心理的自然类,而联结主义心理学却不能。"(同上:111)

取消论

378 如果上述观点正确,那么我们上面考虑的那类联结主义网络,就不具有任何常识命题态度的后三个特征(命题模块):系统是离散的,可语义解释的,以及因果产生的状态。如果我们同时赞同上面的论题[C],那么联结主义网络就不能与命题态度相对应:

(1)如果联结主义是正确的,那么常识心理态度就不是命题模块的(见上述 HA,NKA)。

(2)常识心理态度,如果存在,那么它是命题模块的。

(3)所以,如果联结主义正确,那么:[E]就不存在命题态度(再次提到)。

但是,对于许多研究者来说,命题态度(信念、意愿、意向等)显然是存在的,因而上述论证似乎可作为联结主义的反证:

(4)确实存在命题态度。

(5)因而,联结主义错误。

正如拉姆塞等指出的:"在这些模型中,没有任何东西能与常识心理学的命题态度建立合理的模拟关系"(同上:116)。"如果我们概括的这类联结主义假设是正确的,那么对命题态度的取消论也应该是正确的"(同上:94)。当然,会有很多联结主义和/或常识心理学的支持者,以各种方式对这个论断作出回应。最初的论证由三个步骤组成,回应可针对其中的任何一个步骤。

I. 联结主义与命题模块

与拉姆塞等的观点相反,这一回应试图说明,联结主义模型能够拥有命题态度的后三个关键特征的状态,例如,联结主义模型也可以支持命题态度的模块性,尽管在单元与权值的描述层次上,对其偶然地进行检验对于这一

点并不是很明显。拉姆塞等(1991)认为联结主义会从以下三个方面来论证这个观点：

1.*联结主义*：相信(P命题)是一种特殊的激活**样式**。

拉姆塞等的回应 网络呈现的具体激活样式，其时间过程是短暂的、转瞬即逝的，而相信(以及其他命题态度)是相当稳定的、持久的认知状态，因而相信(以及其他命题态度)不是激活样式。

我们知道，存在现时信念与持久信念的区别，可认为现时信念是一种激活样式。因此这一回应，尽管可认为持久信念不是激活样式，但并不能说现时信念不是激活样式。

2.*联结主义*：相信(P命题)是产生某一特殊激活样式的倾向。

这就避免了时间长度问题，因为这些倾向是可以长期存在的。一种倾向可在网络活动停止后继续存在，就像我们可以相信某些事情，尽管此时此刻并没有思考这些事情。

拉姆塞等的回应 倾向与信念(以及其他命题态度)的持久状态是不同的，因为倾向并不是某种离散因果所需要的激活状态。尤其是，考虑到前面阐述的等势情境，这一点就更加明显了。爱丽丝可能是想到办公室见她的助手，但事实上在具体情境中她查看了邮件，"很难发现，联结主义所讨论的网络倾向，能够处理诸如此类的这些区别……在一个如网络'A'的分布式联结主义系统中，产生某一激活样式的倾向状态，在功能上与产生另一个激活样式的倾向状态是不可分割的"(Ramsey et al., 1991, 115-16)。值得注意的是，这一回应的立论根据是，无法找到常识心理在联结主义网络上的映射关系，并非直接针对这一观点本身，因而与下一个也是最后一个反对回应的关系更为密切。

3.*联结主义*：信念(P命题)仍是一个**有待发现的**系统的特征——命题编码离散系统的一些潜在功能。

拉姆塞等的回应 这当然可能是正确的，但需要有更充分的理由支持，否则这只能看作是对问题本身的回避。

克拉克(Clark, 1990)的策略是想表明，在联结主义网络中，确实存在着与相信(以及其他命题态度)类似的状态，但只是在更高的描述层次上。克拉克说道："分布式的、次符号的，以及包含多个层次的联结主义模型，其组织结构事实上要比拉姆塞等所认为的要复杂得多，因而明显地与命题模块

380 所要求的特征是相容的。因此,我们可能需要质疑,为什么只选择单元一并一权值描述,作为唯一的科学心理网络的描述"(1990:90)。克拉克以NETtalk 为例,尤其是谢诺沃斯基和罗森伯格(Sejnowski and Rosenberg)隐单元层激活的"层级聚类分析",这种隐单元层的激活依据元音和辅音,能够在权值域(weight space)中产生连贯的语义解释。克拉克的基本观点是,(持久的)相信也许可以通过诸如 HCA 等技术手段,能够在足够复杂的系统(网络"A"可能还不够复杂)中实现。他说道,拉姆塞等声称能够"从分布式的、次符号的储存和表征的论证中,可直接推导出取消主义。但从束分析仅有的可能性中,正如我前文指出的,事实上我们并不能直接推导出他们的这一观点"(1990:94)。

II. 对取消论的挑战

以下的回应试图说明,能够引出命题态度非存在的推论并不成立,即联结主义网络缺少"命题模块"的特征,并不是命题态度在系统中不存在的充分条件。我们首先需要明白,对某事物特征的**误判**和表明这个事物**不存在**,是完全不同的两回事。在第 3 章中,我们提及某些著名的古希腊哲人认为脑是血液的散热器,而我们现在认为是心脏;还有一些中世纪的著名学者认为,思维在脑室中发生,我们认为,他们关于脑(和心脏)的理解是**错误的**,而不会说脑和心脏(如他们所设想的)**不存在**。原因可能出自以下理由之一:

1. 允许从对命题模块的错误描述,推导出命题态度不存在的原则是有缺陷的(**原则错误**);

2. 事实上,这三个命题态度的特征并非都是正确的(**常识错误**);

3. 命题态度的常识概念构成不仅只有这三个特征(**常识描述错误**)。

下面,我们逐一进行讨论:

1. 原则错误

这一点是斯蒂克和沃菲尔德(Stich and Warfield, 1995)所作的挑战,反对拉姆塞等能够得出取消论的推论。他们是这样看待这个问题的:联结主义对常识心理学命题态度是**错误理解**的这一事实,与推论出联结主义必然

381 认为命题态度**不存在**的结论,两者之间是无法建立合理连通桥梁的。他们指出并反驳了能推论出取消主义的两个原则。首先,存在:

(DTR)

指称描述理论(description theory of reference):通过满足与每一术语联结的某些(或许是权值)描述集合,理论中包含的术语能够指涉实在。

因而，如果联结主义网络中没有符合命题态度模块性的描述，那么网络不包含它们。

回应：斯蒂克和沃菲尔德回应说，现在的语言哲学主流观点是，描述理论实际上是错误的，而与描述理论相对的“历史一因果链(historical-causal chain)”理论，“能够证明是关于大多数理论术语指称关系的正确解释”(Stich & Warfield, 1995:406)。在这种相对立的解释中，术语通过历史一因果链溯源其所指的实体而获得指称(尽管这一理论的作家们通常很少对心理术语进行讨论)，而且谈话者或思考者并不一定非要知道对于指涉事物的正确描述(或许有人可以谈论恺撒，但他所说的关于恺撒的事情都是错误的)。如果这就是命题态度术语的运作方式，那么在系统中，没有发现能够用命题态度模块术语描述的状态，并不能说明系统就没有命题态度——只能说明命题态度不具有那些描述特征或者我们误解了。然而，“大多数的理论术语”是以这种方式运作的观点仍是有争议的，因为至少还有一些似乎通过描述而被引入的术语。而且，这样的理论术语的运作方式，多数都局限于物理学和生物学领域——无法确保心理学的(心理的)术语也以同样的方式运行。最后，我们还不清楚是否能够把常识心理概念与科学理论术语进行同化。

(CP)

本构特征(constitutive properties)：某些事物的本构特征，就是该事物的特征，或它的本质特征。如果系统不具有这些特征，那么由这些特征所构成的事物也就不存在于这个系统中。

比如，金元素拥有79个质子(元素原子序数79)是它的一种本构特征，如果我们发现某元素有92个质子，我们就知道这是铀而不会错误地认为是金。同样，如果联结主义网络没有任何状态是由作为命题态度模块的三种本构特征所构成，那么系统就不具有常识命题态度。

382

回应：斯蒂克和沃菲尔德回应说，首先，很难确定某些特征是，而另一些特征不是事物的本构特征——仅仅认为它们应当是，这是不够的；第二，奎因(Quine)及其他人对如下观点提出了质疑，即概念整体是否具有本构特征，或者语言模式中是否存在着“分析式地”与指称术语联结的特征——语言所指对象的特征，术语必定能够对其进行指称。传统的解释与我们想要解决的问题背道而驰，“如果有哲学家要选择这条道路继续下去，我们只有祝福他们，但我们并不会真的屏住呼吸等待他

们的成功”(Quine, 1990:409)。但这里,我们或许还需要补充一句,本质特征是否存在依然是有争议的话题,而且在心理学(心理)概念领域对这一问题的研究尚不够充分。

到目前为止,对于取消论我们能够得出哪些结论呢?很明显,通过说明DTR和CP这两个具体的原则,无法从对命题态度的误解推论出它们不存在,但并不能表明没有其他的原则能够做到这一点。拉姆塞等(1991)提出的这两个原则是基于他们的原初论证,而斯蒂克和沃菲尔德却没有对这个原初论证加以考虑,因而就显得有些讽刺了。正如拉姆塞等(1991)指出的,科学哲学还并未形成任何判定误解与不存在的原则,因而从根本上说都只是寻求某种判断的呼吁,但是:

(DFD)

“如果一种新理论的假设,让我们觉得是**深刻地和根本地不同于**(deeply and fundamentally different,DFD)旧的理论……那么就可以近似合理地得出结论,这样的理论变化是根本性的,取消论的推论也就合乎程序了”(同上:96)。

这表明,因为联结主义网络看起来很合理,因而似乎就不包含任何“模块的”命题态度。那么,从取消论中我们又能得出什么结论呢?联结主义模型能否缺少“模块的”状态,但依然具有命题态度?仍然有两种回答需要思考。

2. 常识错误

拉姆塞等(1991)认为,命题模块的这三个特征全部是常识命题态度概念的一部分,而以斯莫琳斯基(Smolensky, 1995)为例,在他所研究的模型中只发现了前两个。斯蒂克和沃菲尔德评论说,如果斯莫琳斯基将来的工作依然无法发现第三个特征,即因果作用,那么他的建构“无法算作是充分回应了拉姆塞等的挑战”。但是,如果有人想要**否认**因果作用要么是(a)命题态度常识概念的一部分,要么是(b)(**无条件地**)属于命题态度的一部分,那么拉姆塞等的挑战就不得不进行修改了。

383

3. 常识描述错误

克拉克(Clark, 1993,继 1991 之后)更青睐于第三种回答,并且建议“彻底抛弃常识心理学必然有助于解释具有直接因果效力的相信和意愿的观点……还存在其他的方式……建构具有解释值的模型——这些方式并不要求与任何具体的基础的科学本质等同”(1993:211)。他对于这一问题采取的策略是,在网络中运用“解释”需求替代因果需求。然后论证就命题态度而

言，对于行为的联结主义**解释**能够满足等势情境。然而，还不清楚常识心理学是否真的不包括特殊的因果作用。

评论

我们上面详细阐述了命题态度在联结主义模型中的地位存在的一些争议。一些论证尝试说明联结主义是错的，因为它并没有像人们普遍理解的那样包含命题态度(只是看起来似乎是包含的)。我们已经看到每种论证及回应都有其各自的优缺点，现在看起来似乎还无法得出确切结论——无论是支持的还是反对的都无法完全获得支持。最核心的问题似乎是命题态度的因果与解释作用的区别，而这一问题还没有定论。

13.6 CCTM 探测器语义

在前面讨论认知科学的神经一逻辑基础时(第 4 章)，我们对于诸如蛙视觉系统的系统形式化地提出了简单探测器语义(SDS)的观点，并且简要讨论了相关的各种问题。现在，我们对于 CCTM，也可相应形式化地提出简单联结主义探测器语义(Simple Connectionist Detector Semantics，SCDS)：

(SCDS)

单元或者单元集表征激活它们(或者能够充分地激活它们)的事物。

在这里，我们继续运用同样的策略讨论，联结主义表征如何应对前面提及的各种问题。正如我们将要讨论的，SCDS 在其**正确性**和**完整性**两方面都存在着严重不足。

真实原因或“纵向”疑难与非原因问题

前面提到的真实原因(right-cause)或“纵向”疑难(“depth” problem)，是指如何在可能的众多因果链中分辨出真正与表征相关的那个因果链。而“非原因难题(no-cause problem)”是指当不存在(相关的)原因时，如何解释表征的语义难题。 384

真实原因

先来看一下麦克莱兰德和鲁梅尔哈特(McClelland and Rumelhart，1981)提出的，著名的关于 1179 个四字母单词的字母识别模型(见图 13.5)。

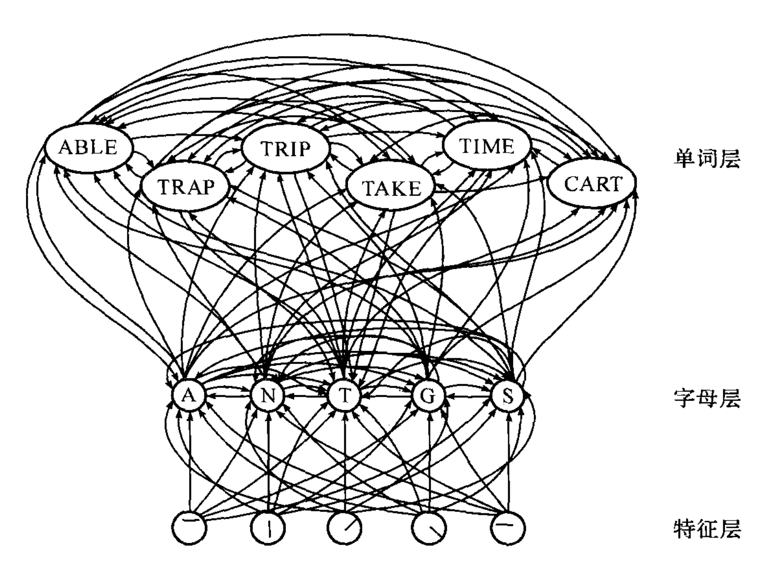


图 13.5 字母识别网络(来源:McClelland and Rumelhart, 1981: 380, 图 3; 经 American Psychological Association 许可复制)

首先,每个单词是一个节点,在不同的位置上组成单词的每一个字母是一个节点,每个字母的特征是一个节点。其次,这些节点被归为三层:特征层、字母层和单词层。最后,节点间的连接存在两种方式的兴奋和抑制。相邻层的节点间具有兴奋和抑制连接,但同一层内的节点间只有抑制连接。这一系统吸引人的地方,不在于其与真实世界的联系(这点很值得思考),正如麦克莱兰德和鲁梅尔哈特所承认的,这一模型"明显绕开了几个有关低层次加工的重要问题"(1981:383),其真正令人关注的地方是系统被相关刺激调整后而产生的行为。麦克莱兰德和鲁梅尔哈特记录下了两件事:第一,"每一特征能否被检测到,具有某种概率 P……[概率将]……随着所呈现视觉性质的变化而变化"(1981:381)。第二,"我们根据系统探测的单元[此时单元处于刺激中]识别节点"(1981:379)。

385

这一模型没有真正的输出节点。输出是依照特殊的规则,将暂存的所有节点的激活样式整合后读取。给予系统**反应强度**:"在反应强度的基础上,在每一次任务循环后,程序能够计算出正确的选择对象被读取的概率和非正确的选择对象被读取的概率"(1981:407)。

字母和单词节点——"隐层"单元如何运作?这些单元作为与它们相连接的每一个节点的激活功能而被激活。以字母"T"为例,它是由聚焦于同一"视网膜"[4]区域的,中央垂直棱边探测器和顶部水平棱边探测器同时激活而激活。因而,"T"节点是因为垂直和水平棱边探测器激活的联合(在时间

和空间上)而激活。单词节点的原理也是这样,只不过需要按位置顺序的四种激活发生联合。单词节点由四个字母一位置节点组激活,例如,“T”在第一位置,“A”在第二,“K”在第三,“E”在第四,就形成了“TAKE”。

(SCDS 的)这些节点都能表征什么激活了它们,而激活它们的就是先前的激活节点。但是我们希望,当然是合理的希望,这个模型能够做到,字母节点能够探测**真实世界**的字母,单词节点探测器能够探测**真实世界**的单词。被识别的应该是**系统之外的**的字母和单词,而不是信息流中先前的节点。模型所依托的理论必须能够找到一种方式处理“激活”的转换——必须找到节点激活的真正原因。这也就是联结主义系统的真实原因难题。

鲁梅尔哈特和麦克莱兰德(1986b)的过去时态学习者模型也是这样,除了其中的 A 单元表征词根(word roots)而 B 单元表征过去时态的形式(past tense forms),两者都属于 Wickelfeature 表征。固定的解码网络(decoding networks)将 A 与 B 的表征,从语音表征(phonological representations)转换过来,并且又转换为语音表征,表征解释很明显是针对听觉的(鲁梅尔哈特和麦克莱兰德没有说明网络如何接受刺激,但最有可能是从键盘输入)。每一个 Wickelfeature 的要素都很复杂,为使计算方便,鲁梅尔哈特和麦克莱兰德在一定程度上对它们重新进行了编排。他们按照“声音”的四个维度进行划分,得到了构成字母的特征的 10 点要素:(1)阻断的辅音 vs. 连续的辅音 vs. 元音;(2)爆破音 vs. 鼻音,摩擦音 vs. 响音,高音 vs. 低音;(3)口腔前部、中部和后部;(4)浊音 vs. 清音,长音 vs. 短音(见图 13.6)。

386

		口腔部位					
		前		中		后	
		V/L	U/S	V/L	U/S	V/L	U/S
阻断辅音	*Stop*	b	p	d	t	g	k
	Nasal	m	—	n	—	N	—
连续辅音	*Fric*	v/D	f/T	z	s	Z/j	S/C
	Liq/SV	w/l	—	r	—	y	h
元音	*High*	E	i	O	^	U	u
	Low	A	e	l	a/α	W	* /o

字母所代表的语音:N=ng in *sing*; D=th in *the*, T=th in *with*; Z=z in *azure*; S=sh in *ship*; C=ch in *chip*; E=ee in *beet*; i=i in *bit*; O=oa in *boat*; A=u in *but* or schwa; U=oo in *boot*; u=oo in *book*; A=ai in *bait*; e=e in *bet*; I=i_e in *kite*; a=a in *bat*; α=a in *father*, W=ow in *cow*; * =aw in *saw*; o=o in *hot*.

图 13.6 Wickelfeatures 的特征(来源:Rumelhart and McClelland, 1986a, vol. 2: 235,表 5;经 MIT Press 许可复制)

现在的问题是：指派辨别这些特征的节点，事实上真的辨别了吗？“浊音”与声带的振动有关，而“前部”、“中部”和“后部”则与口腔发音位置有关，这些节点能辨别口形吗？或者，因果链只需回溯到由口形产生的声音特征（如果有的话）即可？[5]

非原因

前面我们已注意到，真实原因难题的极端情况是“没有原因”难题——我们应该如何表征那些不具有产生表征殊型（tokenings of representations）的东西？这些包括逻辑概念，如 XOR；抽象实体，如数字、集合等；未来实体，如我们尚未出世的子孙后代；可能还包括社会角色，如作为纽约市市长。看一下线性阀阈单元表达 XOR 网络的例子，如图 13.7 所示。在这一模型中，存在一个隐单元（“内在的”），如果没有这一单元，XOR 将无法计算。但是所有这些单元根本没有指派明确的解释，更不用说探测语义特征了。它的意义，更多地在于其在网络中起到的功能作用，而不是表征什么激活了它。

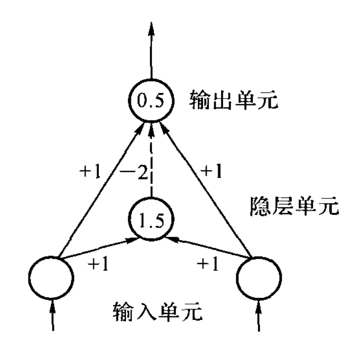


图 13.7　XOR 网络（来源：Rumelhart，1989：151；经 MIT Press 许可复制）

错误表征（析取）难题

387　就像我们讨论“蛙眼告知蛙脑什么”（第 4 章）的问题一样，这个问题也是由福多提出的，即人们所熟知的**“析取难题”**（disjunction problem）。对联结主义系统而言，这个问题可能是这样的：假设作为一个节点的“T”探测器的定义为，只要接触到字母“T”[6]时便获得开启（on）。然而，正像所有的装置偶尔都会出错一样，它也会偶尔地比如接触到“I”时也会激活。这样，现在的 T 探测器在功能上就等同于 I 探测器了，也就是说，它总体上是个“T 或 I 探测器”。因而，似乎看起来错误表征是不可能的了，一个可能的错误表征的例子被理论转换成了纯粹的析取表征问题，这并不是我们所想要说明的。

蕴含难题

另一个联结主义探测器语义特征的问题，是如何确定**表征间**的语义关系。我们在前面看到，近似于人类认知的复杂表征系统，至少具有某些逻辑蕴含关系，而难题是网络如何能够获得这些关系？这就是联结主义系统的蕴含难题(entailment problem)。我们注意到，NETtalk 可以说至少获得了一些原始关系：它能表征"a"是一个元音，其意义是如果一个"a"单元被开启，这一激活会扩散到它的隐"元音单元"并使之激活。但是，正如我们前面在语义网络中看到的，问题是如何控制这种传递仅仅认为是一种推理过程。

"作为什么"的表征，"横向"或"特征"问题

388 对探测器语义表征进行以下两方面的区分很重要。一方面，表征是**对于或关涉**(of or about)**真实世界的什么**的表征。再次回顾第 4 章"蛙眼告知蛙脑什么"。假设一只在自然生境的，功能健全的蛙，具有**对于或关涉昆虫**的表征。也就是说，作为历史事实，在野外是**真实的昆虫**开启了蛙的昆虫探测器。另一方面，表征是如何表征这样的事物——是将事物**作为**(as)什么进行表征。解决了这一问题，也就解决了我们称之为(第 4 章)"横向"或"质性"难题("spread" or "qua" problem)。有人描述其中的一种特征可能是，"很小的事物(3 个单位或者更少)"和一个"苍蝇大小的事物，1 单位大小"。很明显这两种事物都**不一定是昆虫**。在我们看来(应该是可以这么说的)，蛙对于"昆虫"的表征，可能无法与它对"一团小小的黑黑的不规则移动的东西"的表征区分开来。我们可能会知道，系统的表征是关于什么的表征，真实世界中的什么物体它正在追踪，但依然不知道它把它作为什么来表征——**作为什么进行表征**(what it is represented *as*)。而且还有一点我们需要注意，即表征可以因某事物作为具有某种属性，而不是另一属性的事物而描述这个事物，尽管它所有探测到的事物(或者甚至对于所有事物)都既有第一属性特征也有第二属性特征。引用德雷兹克(Dretske，1986a)的一个例子，电梯门上的电子眼可以探测门前是否有事物出现，尽管只有人(或神秘魂魄)才使用电梯，但我们不会把电子眼说成是"人(或神秘魂魄)探测器"[7]。探测器语义特征似乎非常有益于回答系统正在表征什么的问题——表征的是关于什么的表征。但在联结主义系统中，探测器语义如何解释"作为什么而表征(representation-as)"呢？或许在第 11 章玫瑰和山羊

的表征网络中能找到回答这个问题的线索。在那里，我们给每个节点指派了一个心理物理"维度"，如颜色、形状、气味和纹理，然后我们允许不同的激活层表征这些维度中的不同点：红的颜色、椭圆的形状、芬芳的气味、粗糙的纹理等等。于是，这些矢量就不仅仅由（因而是关于）一朵玫瑰或山羊而产生，而是把它**作为**具有红色的、椭圆形的、芬芳的等等特点，也就是**作为玫瑰**来表征。

结 论

这又会把我们引向哪里呢？如果一个系统需要另一个已经具有语义的系统给予它语义，那么我们把该系统的语义称作**派生**。而不需要其他语义系统给予语义的系统语义称为**非派生**。数字计算机典型地具有派生语义；人类的思维普遍来说是非派生语义[8]。我们所处的现状是，联结主义模型的输入层节点几乎都没有非派生语义。这就是说，在联结主义网络中，绝大多数节点或激活样式不具有非派生语义，不论它们得到的什么语义都是派生的——程序员通过符号给予它们。而程序员的表征系统，当然正是联结主义试图模拟的那种认知能力，因而在模型中预设这种能力是行不通的。

389

13.7 CCTM 的问题与前景

CCTM 模型面临着一些与 DCTM 同样的难题，包括意识和感受质性(qualia)难题，但是 CCTM 现在还有各种具体的亟待解决的问题，包括：

1. 它们很难处理瞬间的涉及范围极其广泛的加工过程（比如，语言生成和理解、检测）。

2. 它们很难处理高度结构化的加工过程（比如，语言生成和理解、检测）。

3. 它们具有尺度增大(scaling-up)的困难——小数目单元和窄领域内可执行的工作，并不能顺理成章地扩展到较大的范围领域。

4. 个体节点和矢量语义模糊不清——它们关涉什么，又如何获得这种关涉(aboutness)。

5. 复杂表征语义模糊不清——复杂表达的语义，是怎样与这个表达的组成成分及成分间关系的语义联系在一起，换言之，复杂表征语义是组构性的吗？

注释

[1] 不能忘记模型不可能模拟被模拟物的所有方面——太阳系模型可以是铜制的。

[2] 尽管一个受到改变也会影响其他的变化——知道了属于内华达州的里诺市,位于属于加利福尼亚州的洛杉矶市的西面,同时可能也会改变某人对于加利福尼亚州和内华达州的地理知识。

[3] 事实上,他们将其称之为"联结主义记忆模型",而且通常把它的内容称作"信息"而不是"相信"。然而,如果不是联结主义模型的相信,就与态度问题不相关了。

[4] 如果它们不是"聚焦"于"视网膜"的同一区域,那么同时呈现"|"和"—"也将激活"T"探测器,尽管它们是分开的,并不构成"T"。

[5] 想要回答这些问题是非常困难的,某种程度上对于听觉语音和发声语音的研究尚不够充分。探测器语义部分地属于实证议题——这也是它部分的魅力所在。 390

[6] 短语"只要……时(just when)",其实本身就包含了一种因果观念——属于一种因果理论。

[7] 实际能够探测的事物是无法穷尽其所有表征潜在性的——我们必须认为,它所探测的事物往往在反事实条件中是正确的(这正是难点之一)。

[8] 其性质可用某种公共语言进行思考。在公共语言中,词汇是关于什么的词汇,由语言共同体或语言共同体内的专家们所确定。

【思考题】

大脑

经典联结主义模型与人脑有哪些主要区别?(提示:神经元与单元、化学、几何、学习、尺度)

联结主义的优点

什么是"福多之叉"?

"100步规则"优点是什么?它是非本质的还是属于执行问题?

人们为什么相信它?

福多和派利夏恩如何回应?

我们应怎样评价他们的回应?

"阻抗干扰和损害"优点是什么?它是非本质的还是属于执行问题?

福多和派利夏恩如何回应?

我们应怎样评价他们的回应?

"软约束"优点是什么?它是非本质的还是属于执行问题?

福多和派利夏恩如何回应？
我们应怎样评价他们的回应？
“规则清晰性”优点是什么？它是非本质的还是属于执行问题？
福多和派利夏恩如何回应？
我们应怎样评价他们的回应？
391 “脑一式样模拟”优点是什么？它是非本质的还是属于执行问题？
福多和派利夏恩如何回应？
我们应怎样评价他们的回应？
是否还有一些“优点”福多和派利夏恩没有涉及？举例说明。
福多和派利夏恩面临的“经典结构的两难选择”是什么？
联结主义模型主要针对的问题是什么？

中文体操馆

塞尔的中文体操馆的情境是怎样的？
塞尔的中文体操馆的论证是什么？它有哪些问题？
塞尔的“串行”论证是什么？它有哪些问题？
塞尔的“并行”论证是什么？它有哪些问题？

命题态度

命题态度的“实在论”和“取消论”分别是什么？
DCTM 的命题态度是怎样的？
DCTM 的命题态度有哪四个特征？
常识命题态度有哪四个特征？
什么是态度的“等势情境”？
CCTM 的表征有哪四个特征？
什么是“命题模块性”？
什么是“整体论”论证？
可以用哪两个主要的策略回应这些论证？

探测器语义

392 什么是初级联结主义探测器语义(SCDS)？
对于 SCDS，什么是“真实原因”难题？
对于 SCDS，什么是“非原因”难题？
对于 SCDS，什么是“错误表征(析取)”难题？
对于 SCDS，什么是“蕴含”难题？
对于 SCDS，什么是“作为什么而表征”难题？
联结主义网络能计算 XOR 吗？怎样计算？

问题与前景

CCTM 和 DCTM 共同的难题有哪些?

CCTM 还有哪些不同的难题?

【推荐读物】

CCTM 与人脑

关于神经元的基础知识,见第 3 章的选读文献以及 Arbib(1995)。Kosslyn (1983)的第 2 章是关于联结主义计算与人脑问题较为通俗的介绍。关于神经元和联结主义网络的更多介绍,见 Crick and Asanuma(1986),Schwartz(1988), Churchland (1989; 1990, 第 1.4 节), Copeland (1993b)第 10.5 节, McLeod et al. (1998)第 13 章,以及 Dawson (1998)第 7.1V 章。对人脑功能网络模型的生物实现的近期讨论,见 Rolls and Treves (1998)。联结主义神经科学的近期讨论,见 Hanson (1999)。

联结主义的优点

对联结主义"优点"的最有影响的论述,见 Rumelhart and McClelland (1986a) 第 1 章。在 Bechtel and Abrahamsen (1991)第 2 章及其他几章也有所讨论。关于联结主义的实现与认知的讨论见 Chater and Oaksford (1990)。McLaughlin (1993),McLaughlin and Warfield (1994)为经典结构进行了一些新的辩护。

中文体操馆

Churchland and Churchland (1991)对中文体操馆论证作了些许回应,更多的讨论见 Copeland (1993a; 1993b, 第 10.6 节)和 Dennett (1991)。Smolensky (1988a)和 Franklin and Garzon (1991)都认为,"严格地说,神经网络对于解释心灵可能要比图灵有力,因而能解决图灵机所不能解决的问题"。我们将在结语部分再回到这个议题。 393

命题态度

本书对于 CCTM 和取消主义的介绍,主要关注了 Ramsey, Stich, & Garon (1991), Clark (1990; 1993, 第 10 章)以及 Stich and Warfield (1995)的相关讨论。更深入的讨论见 Smolensky (1995),它对 Ramsey, Stich, and Garon (1991)以及命题模块性作了回应。我们的讨论更多受益于 Braddon-Mitchell and Jackson (1996)的第 3 章和第 14 章,以及 Rey(1997)的第 7 章。关于常识心理学取消论的最有力讨论见 Churchland (1981)。关于常识心理学问题的简要介绍,见 von Eckhardt(1994)和 Churchland(1994)及其参考文献。最近的著作见 Greenwoood (1991), Christensen and Turner (1993), Stich (1996)第 3 章以及 Carruthers and Smith (1996)。

探测器语义

关于“析取”难题的更多讨论，见 Fodor (1984，1987，1990)。关于“作为什么而表征”见 Dretske (1986a) 及其参考文献。对 SCDS 的讨论，见 Ramsey (1992)。

问题与前景

对早期 CCTM 议题的回顾，见 Rumelhart and McClelland (1986a)第 1 章和第 26 章。对联结主义困境和前景的总体讨论，见 Rumelhart and McClelland (1986a)第 4 章，Minsky and Papert(1969)及其 1988 年重印版的后记，Rosenberg (1990a，b)以及 Quinlan (1991)第 6 章。关于联结主义一些较早的实证批评，参见 Massaro (1988)以及 Ratcliff (1990)。

结语:认知科学的计算或者究竟什么是计算机?

C.1 引 言

我们一直假定心智的计算机模型是可行的。从整体上或直观上来说,394 计算机可以看作是这样一种装置,它包含一种能够对(关于世界的)表征进行操作的程序结构。确切地说,就是心智计算理论(CTM):

(CTM)

(a)认知状态是指具有内容的心理表征之间的计算关系。

(b)认知过程(认知状态的变换)是对具有内容的心理表征进行的计算和操作。

在第二部分,我们详细阐述了心智的数字计算理论(DCTM),就是在CTM计算方式中加入了数字限制。在第三部分,我们阐述了联结主义的心智计算理论(CCTM),是在CTM的计算方式中加入了联结限制。

现在我将阐述一种关于计算机和计算更普遍的观点——尽可能涵盖所有现实机械装置。假设有一种机器突然出现在面前,我们想知道它是否一台计算机——是否能够进行计算,那么我们需要知道它什么呢?有两种宽泛的理论可用于理解计算机和计算的特征——一种是从装置所具备的**功能**,另一种是从装置的**描述层次**。每种方式都存在一定的优点和缺陷,我们将简明审视这两种理论,然后尝试是否能将两个理论合成一个更有效的观点。

395

术语解注 我们反复使用"计算机"和"计算"这两个概念，这在前文中并不常见。例如，字典中经常用计算来定义计算机，反之亦然，"计算机是一种能够计算的东西（见下）"和"计算就是计算机所做的事情"。明显这样定义是并不充分的，其中：(1)计算机能做的不仅只有计算（它还可以散热，指示灯不停地闪烁等）；(2)人也能够进行计算，但人并不是计算机（即使人的思维可能是计算的）；最后(3)两个概念循环定义。我们不得不在某处打破这种循环，虽然本应该是在"计算"的概念上有所突破，但我们将重新考察"计算机"的概念。

C.2 计算机功能主义观

考察计算机和计算的定义，一种方法是从已有的理解开始，主要有两个来源：字典和教科书。

字典的定义

一个具有代表性的示例：

> 计算机
> 1. 能够进行计算的某人或某物。[按算法进行的计算]
> 2. 能够高速执行重复和高度复杂数学运算的机器或电子设备。
> （兰登书屋英语词典，完整版：第 303 页）

这两种定义存在着很多问题：第一种定义，依然是上面提到的循环定义；第二种定义，它排除了很多或者所有具有计算能力的人。只要想一想一个不会算术的，或者丝毫不会重复复杂数学运算，更不用说是高速运算的人，会是什么样子。

教科书

（别期望能清楚地理解这个定义）[1]

定义 3.1 程序 P 在机器 M 上的一个（完整的）计算是一个有限序列：$L_0, m_0, L_1, m_1, \cdots, L_n, m_n$ 分别对应 P 的转换字符和 M 的要素。L_0 是 P 的起始指令，L_n 是 P 的中止指令。当 $i<n$ 时，要么获得形式指令 L_i：执行 F；切换到指令 L'，$L' \in P$，$L_{i+1}=L'$ 对应 $m_{F(mi)}$，要么获得形式指

396

令 L_i：如果 P，那么切换到 L′，否则切换至 L″，L″∈P。如果 $m_{i+1}=m_i$，要么 $m_{p(mi)}$ = T 且 L_{i+1} = L′，要么 $m_{p(mi)}$ = F 且 L_{i+1} = L″。（Scott 1967：193）

这个定义存在的问题是它需要依赖先前对程序和机器的定义，事实上，"计算"也是用这些术语定义的（这样就使得计算机运行计算等价于 UTM）。

纽厄尔：物理符号系统

经常提及的一个计算机概念是纽厄尔（Newell，1980；Newell and Simon，1976）提出的，称作"（物理）符号系统"："一种宽泛的系统类别，能够持有并操作符号，但在物理世界依然是可实现的……"（1980：38）。（物理）符号系统包含五个子系统：存储、操作控制、输入和输出（见图 C.1）。

397

SS：符号系统示例

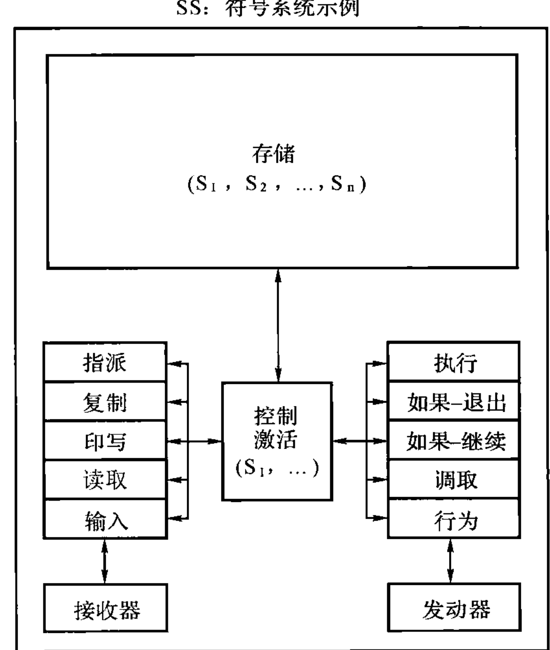


图 C.1　符号系统结构范例（来源：Newell，1980：45，图 2；图 C.1—C.3 经 *Cognitive Science* 许可复制）

存储系统由一种特殊类型的"符号结构"或"表达式"构成，起到一种特殊的作用。系统有 10 种操作（指派、复制等），每一种操作负责符号的输入以及符号的输出，这个过程就是"控制"（见图 C.2）。

指派 S1,S2 到同一物理组织(entity) 　　得到带有新指派的 S1	(指派 S_1S_2)
复制表达式 E(产生新的符号) 　　得到新的表达式和符号	(复制 E)
印写 S1 到表达式 E 中作用 R1 　　得到修正的表达式 　　零变动不存在	(印写 ER_1S_1)
读取 E 中的作用 R 　　得到表达式或维持原貌	(读取 R, E)
执行 S1,S2,S3…序列 　　得到由上一个 S1 产生的表达式	(执行 S_1S_2)
如果上一步得到结果 E 仍是原来的 E,那么退出序列 　　得到原来的 E	(如果 E,则退出)
调取符号 S 　　得到无解释的 S	(调取 S)
根据表达式 E,产生外部行为反应 　　得到反馈表达式	(行为 E)
根据表达式 E 输入 　　得到新的表达式或零	(输入 E)

图 C.2　符号系统的操作(来源:Newell, 1980: 46, 图 3)

控制系统就是通过启动一种操作;使相应的符号呈现激活状态,然后执行(转译)对这些符号的操作来控制系统的行为,接着开始准备下一次操作。机器如何进行控制见图 C.3。

转译激活表达式:

　　如果表达式不是一个程序:
　　　　结果就是表达式自身。

　　如果表达式是一个程序:
　　　　转译每个符号所对应的角色作用;
　　　　然后执行在这些输入的操作;
　　　　执行操作后的结果就是最终结果。

转译结果:

　　如果结果是一个新的表达式:
　　　　转译使它具有表达式同样的作用。

　　如果结果不是一个新的表达式:
　　　　使它产生符号的作用。

图 C.3　符号系统的控制操作(来源:Newell, 1980: 48, 图 4)

从物理符号系统等价于普适图灵机这一角度上说,这种机器也是普适的:"如果适当的指令通过输入系统,(它们)就可以有与其他机器相同的行

为……它们能够产生其他任何不论如何定义的机器的所有输入一输出函数。

对这种机器，纽厄尔又进一步说明他为什么将其称之为“物理符号系统假设”：“物理系统，具备普遍智能行为的充分必要条件是它需要是物理符号系统”(1980:72)。(这有些超出我们的目的了——我们是为了要知道什么是计算机。)

当然，(物理)符号系统还存在一些困难的地方，很多机器并不是这种普适机，包括所有各种特殊用途的图灵机，以及大多数袖珍计算器。还有，即使允许机器中具备所有的普适操作集，整个组织系统也与冯·诺依曼机很相似，它看起来只是在结构上很特殊因而不具备计算机的普遍特征。

冯·艾克哈特

另一个更笼统但贴近常识的定义由冯·艾克哈特(von Eckhardt)提出：“计算机是一种凭借对信息的表征进行输入、存储、操作和输出，从而能够自动输入、存储、操作和输出信息的装置。信息加工按照有效的，某种意义上存在于机器自身之中的有限规则集而得以实现”(1993:144)。比起第一种定义具有很大的改进，因为它认识到那些精密的数字游戏对计算而言并不是本质的，而且还注意到了程序和表征所起到的作用。不过它使用的概念，如“存储”和“操作”信息，还需要进一步说明，但也较容易，如：一般而言，纽厄尔对计算机的定义太过具体了。而且，还有一个特征需要明确说明，“规则”并不一定需要作为一种程序清楚的储存在机器的记忆系统中，我们把这个问题留在 C.4 节中讨论。总的说来，这种机器装置的特征也许跟人类很相似，它也许是最接近心智**标准数字**计算模型的一种具体体现。但还是太过具体了，我们现在已经了解了一些心智理论的内容，所以需要一种更全面的理解，不仅仅是计算机是什么——也许与人很相似——而是什么才是**完全的**计算机?

C.3 计算机描述层次观

到目前为止，我们一直试图从计算机的抽象概念或概括现有机器的细节来讨论，但我们似乎仍不能穷尽所有的特殊结构，让我们尝试另一种策略。 399

福　多

福多(Fodor)提出了一种更抽象的计算机概念:"作为计算机系统,只要它能够使物理设备状态与计算语言公式相匹配,建立维持这些公式间所需语义关系的映射……这样一种考虑是可行且合理的。例如,我们可以以这样的方式指派一组机器的物理状态对应一组语言句子:如果 $S_1 \cdots S_n$ 是机器状态,$F_1 \cdots F_n$ 是与 $S_1 \cdots S_{n-1}$,S_n 分别对应语句,那么只要 F1…Fn+1 可作为 Fn 成立的前提,具有如此物理构成的机器就能运行其状态序列"(1975:73)。既然这个定义更加抽象,那它要说明什么呢?首先,福多描述的是一种计算的特征,而不是计算机,那么我们只好说,计算机如福多讲的是能够进行计算的某种东西。就是说,把机器的某些状态(S)可以等同于表征,通过给这些机器状态指派句子或者"公式"(F)就可以产生表征特征(机器的"思维语言"):

$$\begin{array}{c} S_1 \\ | \\ F_1 \end{array}$$

400　机器从前一表征状态到下一表征状态的移动:

$$\begin{array}{ccccc} S_1 & \rightarrow & S_2 & \rightarrow & S_n \\ | & & & & \\ F_1 & & & & \end{array}$$

每一个表征状态都指派一个公式:

$$\begin{array}{ccccc} S_1 & \rightarrow & S_2 & \rightarrow & S_n \\ | & & | & & | \\ F_1 & \rightarrow & F_2 & \rightarrow & F_n \\ \text{“P”} & & \text{“P}\rightarrow\text{Q”} & & \text{“Q”} \end{array}$$

这种状态序列构成了一个从"P"到"Q"的"证明":"如果 P,则 Q。"但这个提议并没有解释机器状态的转换,只是提及公式之间的语义关系,即如果"P"和"P→Q"为真,那么"Q"必为真。输出始终维持与公式之间所需的语义关系,福多也没有说为什么这种语义关系,需要用计算机概念的一部分——尤其是他用了"证明"来进行解释。这根本就不是一个语义概念,而是一个句法(例如,公式字符串是一个公理系统的证明,从一个公理和其后对应的公式开始,结果或者是公理,或者还是对公理应用的推理规则,或者是先前派生的公式)。

马　尔

马尔(Marr)提出计算机(或者他所说的"信息加工装置")可以有三种不同层次的理解:"[最基本的]只有理解了这三个不同层次,才能说完全理解了信息加工装置"(1982:24)。马尔接着说:"对这三种层次的描述,都将使我们最终了解知觉的信息加工,当然它们是逻辑和因果相关的……还有一些现象用其中的一个或两个层次就可以解释"(1982:25)。

> *计算理论*:计算的目标是什么?为什么它是适当的?通过什么样的逻辑策略以及能用来能执行什么?
>
> *表征与算法*:这种计算如何能够执行?尤其是输入的表征是什么,以及这种转换算法是什么?　401
>
> *硬件执行*:这些表征和算法如何实现?

按照马尔(他讲的主要是分析层次,而不是描述层次)的观点,装置需要独立于公式间的语义关系,在"硬件"和"算法"层次上能够被描述,我们才可把它称为计算机。就是说,从它正在做的和为什么这么做的转换角度,描述机器的行为。然而,"计算理论"层次是否是计算机概念的一部分还并不明确,如果在一个特殊装置中发现了计算的特征,那么就能部分地说明这不是一个好的策略。

丹尼特(Dennett,1991: 276)把马尔的层次观点与他自己的"视角(stances)"理论结合在一起:**意向视角**、**设计视角**以及**物理视角**(尽管与派利夏恩提出的三个层次相结合会更好,见后文)。这些视角潜在的观点是,一般而言,说是采取某种视角能够对系统或丹尼特所讲的"对象"提供预测和解释的工具。特别是采用对象的**物理视角**,需要确定"它的物理构成(也许会一直指向微观物理层次)及其所产生的物理性质,以及使用物理规律的知识预测任何输入的结果……这一策略并不总是实际有效的,但它在**原则上**坚持物理科学的信条"(1987a:16)。这相当于马尔的**硬件实现**层次。采用对象的**设计视角**,"可能会忽略(可能是混乱的)对象物理构成的实际细节,并假设它有某一设计,预测它在各种情况下的**行为遵循这种设计**……只有按设计执行的行为是可以预测的,从设计视角而言,当然如此"(同上:16-17)。这相当于马尔的**算法和表征**层次。最后,采用对象的**意向视角**,首先第一个近似是,"把布伦塔诺(Brentano)和其他人所说的**意向性**作为对象的组成部分,将对象看作是具有相信和其他心理状态的理性实体,进而预测它的行

为”(同上:15)。第二个近似是,“首先需要把对象看作是理性实体,它的行为能够预测;然后,找出实体应该具有哪些信念,给出它在世界中的位置和目的。出于同样的考虑,接着需要找出它应该具有哪些欲望。最后根据理性实体的相信所产生的目标,预测它的下一步行动”(同上:17)。这相当于马尔的**计算理论**层次。

402

派利夏恩

计算机是独具三种描述层次的装置,这一观点被派利夏恩(Pylyshyn, 1989: 57)称为“经典图视”,虽然他对三个层次作了少许修改。按照派利夏恩的说法,计算机(还有心灵,如果心灵是计算机的话)需要至少具有下面三个不同的组织层次:

1. *语义层次(知识层次)* 在这个层次上,可以解释为什么适当编程的计算机,通过它所知道的东西,以及与有意义的乃至理性的方式所进行的联结而具有的目标,能够完成一些事情。

2. *符号层次* 知识的语义内容和目标,假定能够用符号的表达形式编码。这种结构表达式有很多部分,每一部分负责编码特定的语义内容。编码、结构以及对它们的操作规则,是另一种系统组织层次。

3. *物理层次* 整个系统运行,必须以某种物理形式实现。结构以及产生物理对象功能的原则对应物理层次。

这在一定程度上正符合福多所关注的,计算机是符号系统,它的**物理**描述和**语义**描述紧密关联(层次 2 和层次 3)。同时也有马尔所关心的,信息处理器需要具有系统的**目标**和它的**合理性**这一描述层次。但是,第二个层次的确切特征却是模糊不清的,例如,“它们[结构表达式]的操作规则”是什么?是马尔的算法层次吗?

福多-马尔-派利夏恩

我们已经介绍了三种观点,以及对它们的评论,我们可以得出一个修正了的计算机“描述层次”概念。希望它至少包含三个层次的描述:**物理**层次(硬件)、**结构**(形式句法)层次(算法)和**语义**层次。如果可能的话,我们也希望能够从装置的**目标**和**合理性**策略方面来评估,最后得到:

403

(F-M-P)最好修正为,如果 X 是计算机,当且仅当它充分满足:

(a)*物理描述*，按照物理法则从一种物理状态运行到另一种物理状态，并与(b)相联系；

(b)*结构描述*，按照普遍结构原则从一个状态运行到另一个状态；

(c)*语义描述*，结构上特定的状态，是语义上对某事的解释；

(d)*合理性描述*（可选)，语义解释的状态负责合理性和连贯性的连接。

我们对层次(c)的描述，是从层次(b)结构(形式，句法)的某些状态的语义或表征特征中分离出来的，目的是得出层次(d)——可选。也就是说，在“非理性”的计算装置中并没有任何固有的矛盾，虽然设计这样的一个装置还没有明确的方法。这种表述意在同时适合认知科学的“经典”框架和联结主义框架，因为暗含(b)中的普遍结构原则可能是硬布线(hardwiring)，或“经典”机器的程序，也或者是联结主义机器中的激活通行和联结强度的规则。

有人担心这种层次描述的观点，完全忽略了提及装置**能够做什么**，因为我们对计算机的概念中似乎有它们能够进行计算的观念，似乎至少也要涉及这种基本的处理能力。但是，在这些人眼中，对描述层次的担心(见第9章)主要还是受功能主义观点的影响。比较而言，功能主义观点的缺点是它过于狭隘，而描述层次的缺点是它又过于宽泛——计算机的定义失败在哪里呢？因此看来，除非我们能够建立严格的“最佳描述”概念，我们也许可以允许行星是计算机，它们能够计算自己的轨道，或(有指针的)手表是计算机，它们能够计算时间。我们稍后会回到这个问题。

C.4　计算机的功能一描述结合观

对上面每一类型计算机定义的忧虑也许是可以解决的，即把它们各自的优点结合在一起。这种结合就形成了计算机的“功能一描述”结合观点。404 其特征如下：

(F-D)　X是**计算机**，当且仅当

1. X能够：

(a)自动进行输入、储存、操作和输出信息，凭借(b)

(b)对信息表征的输入、储存、操作和输出；

(c)这些信息加工过程，在某种意义上就是机器按照自身的有限规则集运行。

2. X 还有一个最佳的描述：

(a)***物理描述***，按照物理法则从一种物理状态运行到另一种物理状态，并与(b)相联系；

(b)***结构描述***，按照普遍结构原则从一种状态运行到另一种状态；

(c)***语义描述***，结构上特定的状态，是在语义上对某事物的解释；

(d)***合理性描述***（可选），语义解释的状态负责合理性和连贯性的连接。

注意到，这个关于计算机的界定方法有这样的优点，它不要求或禁止任何特殊**类别的材料**，但只需要材料有充分的因果条件，使机器能够从一种状态运行到另一种状态。这个界定也不需要机器有任何特殊**框架**，或对表征进行任何特殊**操作**，完全依靠它的一般物理性质。如果我们理解非物理因果关系可能会是什么，我们需要更概括地重新表述条件(a)：

2(a′) ***因果描述***，按照因果法则从一个状态运行到另一个状态。

我们现在将详细说明这一概念，区分计算的不同种类和层次，如硬布线与程序控制。

C.5 计算层次：斯特布勒

根据斯特布勒(Stabler，1983)的观点，计算装置普遍需要具备多种层次或种类，每一种层次或种类在某种意义上决定了(F-D)(1c)是否为真：

405

1. 系统仅需要**计算**函数 F：这里所说的只是指系统能对每个给定输入产生正确输出。

2. 系统**执行**(硬线连接)**程序** P，程序 P 计算函数 F：通过一些中间步骤，以某种方式从输入获得输出。

3. 系统**应用程序** P，支配 P 的执行，计算函数 F：这里重要的事情是在接下来的步骤中控制机器其他的程序状态。

从层次 3 到层次 1 构成一个“标准结构(isa)”等级(见第 7 章)，也就是说任何系统在层次 3 后面是层次 2，层次 2 后面是层次 1，而不能相反。下面，我们阐述这些层次更详细的特征：

层次 1

系统 S 能够**计算函数** F，当且仅当 S 能够实现这样的功能：如当且仅当，

系统的物理状态和某一符号系列存在 1-1 的映射或编码(一种“实现函数”)关系。这样在适当的条件下(机器没有故障等),系统将一直凭借应用的因果法则,从一个起始状态(S_i)到另一个状态,再到结束状态(S_f)。对于每一对这样的状态,与结束状态(S_f)联结的符号就是与起始状态(S_i)联结符号的函数 F。(这里似乎就是福多提出的概念)

在这个概念中,系统能够计算一个函数(例如,计算后继函数,从数字 n 到后继自然数 n+1)。系统状态与符号或者公式 F_1,F_2…(如,0,1…)一一对应,这样当系统处于 S_i 状态(如“23”)时,这个状态通过实现映射与某一符号 F_u(如“23”)相联结,然后机器就会进入 S_f 状态(如“24”),这个状态也与实现映射符号 F_{24}联结。

斯特布勒(Stabler, 1983:402)提出,因为与层次 1 的计算描述相应的奇怪映射允许任何(相应复杂的)系统实现任何计算,这样实在**过于笼统**了——什么东西都可以是一台计算机。例如,想一下我们有一套公式,能正确描述行星的运行,也就是说,如果行星在时刻 t_1 位于 L_1,它会在时刻 t_2 位于 L_2 等等,那么这也就是行星时间和位置方程解的 1-1 映射。行星真的可以计算它的轨道!? 这完全不是我们想要的。他建议的解决方法是,“一定 406 要绝对限制映射的实现”。一种可能是严格要求“实现”关系,规定物理系统的一些状态需要借助任何函数“F”符号表达。很难理解这样的限制究竟是什么,但如果主张认知是一系列计算,那就非常重要了。但我们并不认为认知是计算,因为所有事物都是计算,我们会在本章最后一节再讨论这个问题。

层次 2

系统 S **执行程序** P(用于**计算函数** F),当且仅当(i)P 包含一组连续的指令 I,…I_n,对应连续函数 FI,…FI_n,这些连续函数是 F 的组成部分;(ii)S 通过连续函数 FI,…FI_n 计算 F。层次 2 是说,程序中的连续指令映射系统的状态转换,一个袖珍计算器也许就是这样的一个例子。

层次 3

系统 S **支配程序** P,**控制** P **的执行**(**用于计算函数** F),当且仅当(i)存在一个 1-1 的实现函数程序,使 P 的指令映射 S 状态;(ii)存在一组控制状态,使 S 能够计算 FI_i,因为确定 FI_i 的控制状态是计算的。如果 S 是处于另一

种控制状态，那么，S也会是计算另一种不同的FI。

在层次2和层次3中，从“运行程序”的概念中分化出另外两个重要的不同概念：**依照**程序操作（层次2）vs.被程序**控制**（层次3）。在层次3中，一个系统必须不仅要符合程序计算步骤的描述（指令），而且在后续的加工步骤中，这些描述自身能够实现**反事实条件控制**。所以，寻找一些（非怪异的，见上）映射，用程序来描述（层次2）对象，要比找到一种（非怪异的，见上）映射，使任意的物理系统能够运行具体的程序（层次3）要更加困难。很明显，按照斯特布勒的解释，程序（软件）与**被编程的机器**硬件并没有区别：“这个术语407（软件层次处于硬件层次‘之上’）是不合适的，尽管它暗示软件是另一事物，而绝非‘硬件’，是另外一种东西而非物理部分或系统的特征。说明这个口语词汇并不正确，现在的这种解释能够清楚地表明，任何程序都是实际物理的和有效因果法则的实现，能够被任何物理系统使用”（1983：393）。

C.6 数字与联结主义计算机

前面我们概述了两种计算机的概念[2]，一种是**计算机能够做什么**，另一种是**装置的描述层次**。然后我们提出了一种**组合的观点**，依照针对程序的计算和使用程序的计算，最终区分了三种类别或者层次的计算——函数计算。有时人们说，对于认知的计算模型而言，联结主义模式会是**另一种选择**，但只有认为“计算”是“数字的”、“连续的”，或者说是“冯·诺依曼式的”时才如此。“功能的”和“描述的”联结主义机器都会是计算机。

从功能上讲，我们归纳计算机的特征是作为**硬件**，能够依据**系统内的规则**进行**输入**、**储存**、**“操作”**[3]（计算）和**输出**信息。联结主义机器与传统数字计算机的特征有某些相似之处：节点的网络结构和联结是硬件结构，激活一传递规则和联结权值是内部规则，输入矢量是它的输入，输出矢量是它的输出，信息储存在连接权值中。它通过把输入与它的连接权值相乘进行操作（计算），按照激活一传递规则传递激活。CCTM是一个程序概念，可以解释为使用一系列的指令控制某些计算过程，DCTM的概念可能在CCTM中很难重构，其他多数心智理论也是如此。注意到我们用DCTM把被运行的程序与程序的算法编码区分开了（两种不同的程序可以用同一种算法编码）。但是CCTM权值的变化和/或激活一传递规则的变化，用来做什么以及如何做呢？这个问题在CCTM的具体算法中（算法的变化，变化了多少？）还没有得到解决。

我们叙述性地概括了计算机包含三个部分的描述：**符号**、**结构**和**物理**层次。显然，联结主义机器也能够被描述为这三个层次：接收语义解释的激活样式（矢量）是符号的，节点和连接系统是结构的，制作以及决定激活一传递属性和它的连接权值属性（如，突触的化学性质或神经元的直径）的材料是物理的。我们将这些特征概括如下： 408

功　能

硬件	节点和连接
输入	输入矢量
存储	连接权值
操作/计算	矢量乘法
输出	输出矢量
系统内的规则	激活一传递规则，连接权值与矢量相乘

描　述

符号	矢量
结构	激活一传递规则，连接权值与矢量相乘
物理	节点和连接

数字计算机：联结主义计算机的特例？

考虑到这种相似性，把数字计算机看成是联结主义机器的一个非常特别的、“不自然”的特例，这种观点并不奇怪：它的联结限制在每单元四个，每个单元只传递两个值：0，1 等等。就是说，我们是从麦卡洛克-皮茨的“神经元”建立起来的一个系统，通过不让“神经元”继续做任何事情以及严格限制它们的联结，从联结主义的节点中建立起麦卡洛克-皮茨的“神经元”。例如：

连接：	限制 4 个
激活传递：	限制值为 0，1
逻辑阈：	限制规则为布尔函数：与，非等

神经网络与图灵机

409 与图灵机相比，有关神经网络计算能力的所有问题，到现在为止还并不清楚。霍尔尼克等（Hornik et al, 1989）提出，某些联结机器（三层前馈）的计算能力接近于图灵机；另一方面，仍有很多联结机器，它们的计算能力还不清楚。当然，这依然是一个开放式问题，是否存在一些“算法”，联结机器能够计算，而图灵机则不能。弗兰克林和卡荣（Franklin and Garzon, 1991）对图灵机与联结机器的关系提出了一个有趣的看法：

> 1. 用联结主义机器的观点来看图灵机意味着什么，他们给出了一个谨慎的特征描述。
>
> 2. 他们展示了一个有效的步骤，以建立一个（同步的二进制线性阈）网络，这个网络具有与给定图灵机相同的表现。

考虑到丘奇-图灵论题（见第 6 章），任何计算函数图灵机都能计算，可以说“任何计算函数都能够通过适当的神经网络进行计算”。

> 3. 他们证明网络不能处理的“稳定性难题（stability problem）”（使一个任意的网络对任意的输入稳定），与图灵证明的图灵机不能处理的“停机问题”类似。

最后，西格尔曼和桑塔格（Siegelmann and Sontag, 1992）证明“使用按顺序（如线性）和合理权值的网络……能模拟所有图灵机”（1992:440）。他们继续注意到，“定理 1 的结果是得出普适加工网络的存在，能够接收部分递归计算函数和输入串的编码描述，完成任何图灵机基于输入串所能做的一切”（1992:448）。西格尔曼和桑塔格继续考察了能够处理 4 个字母和 7 种控制状态的特定明斯基通用图灵机（见第 6 章），存在着一种“带有 1058 个处理步骤的通用网络……而且很有可能经过仔细构建能够完全还原这种过程”（1992:449）。基利安和西格尔曼（Kilian and Siegelmann, 1993）总结了西格
410 尔曼和桑塔格应用特殊的、很少使用的传递规则构建的网络特征，在很多三层反馈网络中都通用的“S 状”规则，如 NETtalk（见第 10 章）：“表明……存在一种普适结构，可以计算任何递归函数……这个技术能够应用于更广泛的‘类 S 状（sigmoid-like）’激活功能，说明图灵机的普适特征，相对而言，也是循环神经网络模型的共有特征”（1993:137）。这就是我们逐渐得到的，最终对联结主义模式的计算能力更为清晰的认识，但很有可能与我们对传统计算能力的理解相违背。

C.7 所有事物都是计算机吗?

正如斯特布勒所注意到的,为了避免任何事物都能够称之为计算机,所以计算机的实现需要有严格的限制条件,但这些要求过于琐碎,以至于认知仅被认作是计算的一种类型。在前面(见第 6 章),我们回顾了塞尔的"句法论证",他主张大脑从任何重要意义上讲,都不会是一台计算机(对于支持的人,他称为"认知主义")。他认为如果所有的事物都是计算机,那么大脑也是计算机,但这是没有意义和价值的。而我们认为,产生争论的部分原因是,对计算和计算机的描述带有偏差。我们依据现在对计算机概念的理解,以及斯特布勒的层次观点,可以清楚地看到,的确有一个地方出错了。按照塞尔的观点,一个对象能够称之为计算机,至少包含层次 2 的描述——如,可能被描述为计算某些函数。如我们注意到的,要想获得关于层次 2 的计算机前理论(pre-theoretic)概念,那么从符号到状态的映射都必须要加以限制。在这里,塞尔的观点可以看作是对这种观点最为强烈的反对,并且对这样的限制是否能够找到也暗含着否定。然而,如果我们至少还对现有的计算机概念有兴趣,并且认识到所有的认知功能(程序)并非硬线连接,那么就能得到计算机需要至少 3 个层次,或者 4 个层次的装置(能够学习算法的装置)。以 wall-WordStar 为例,如果墙能够"解释"为运行的环境,那么,墙被描述为 WordStar 的实现,就需要满足墙是计算机。但并不能得到这样的结果,层次等级是从层次 3 到层次 2 再到层次 1,而不能相反——例如,有些东西可以用层次 2 描述,但不能够因此而满足层次 3 描述。沿着这个问题我们得出,当程序在机器上运行时(层次 3),机器的程序状态与机器的控制状态序列可能是反事实条件的。塞尔并没有说明墙还具有这种特征——只是宣称墙可以用程序步骤(层次 2,执行程序)描述。应该注意,斯特布勒的等级层次只能应用于这样的条件(2b)——但计算机的所有其他条件,尤其是(2c),是如何满足的呢?墙是否真的获得了描述的语义层次?看起来似乎还没有。 411

注释

[1] 这个定义非常典型,并且被反复提到,例如 Clark and Cowell (1976)。

[2] 我们还要注意,或者反对,纽厄尔将计算机的定义等同于普适图灵机。

[3] 注意到,"操作"符号并不意味着对它们的运行,也不是使用储存于计算机内存中的"规则"操作它们。它只是涉及生成、更改或者删除符号,以及联结主义加工对这三种情形所能够做的一切。

【思考题】

计算机的“功能”观点是什么？
纽厄尔的“物理符号系统”是什么？
冯·艾克哈特的计算机概念是什么？
计算机一般“层次描述”观点是什么？
福多的计算机概念是什么？
马尔关于信息加工装置的三个层次描述是什么？
马尔的“层次”观点与丹尼特的“视角”观点的关系如何？
派利夏恩的计算机概念是什么？
我们怎样把福多、马尔和派利夏恩的观点结合在一起？
计算机的功能－描述组合观点是什么？
斯特布勒的三个计算层次是什么？
为什么说联结主义机器是计算机？
所有的事物都是计算机吗？为什么是，或为什么不是？

【推荐读物】

412 结语中的评论性材料主要来自关于计算的非专业性文献。数学、逻辑和计算机科学方面的专业性文献不在讨论之内。

计算

有关物理符号系统，见 Newell and Simon(1972)，以及 Newell and Simon (1976)。Von Eckhardt (1993)第 3 章“计算假设”，Glymour (1992)第 12 章“可计算”，是针对计算的很好的非数学性导论，类似的还可见 Copeland (1996a)。更专业的讨论可见 Minsky (1967)，Davies (1958)，以及 Boolos and Jeffrey (1989)，其中 Davies (1958)是经典教材，Boolos and Jeffrey (1989)讨论了计算与逻辑的关系。目前有很多关于计算的优秀教材，其中 Clark and Cowell (1976)曾在认知科学研究中(如 Stabler, 1983)被引用。Odifreddi (1989)第 1 章，对计算和递归算法作了很好的长篇(超过 100 页)总结。Hornik et al. (1989)提出了多重反馈网络与图灵机近似的论证，Franklin and Garzon (1991)说明了“任何计算函数都能够通过适当的神经网络进行计算”。

所有事物都是计算机吗？

Putnam (1988)和 Searle(1990b,1992,第 9 章)重点讨论了当前对计算机的标准定义使得所有事物都可能成为计算机的观点。Goel (1992)，Chalmers (1996a)，Chalmers (1996b)第 9 章，以及 Harnish (1996)对此都作了进一步讨论。

参考文献

Akmajian, A., Demers, R., Farmer, A., and Harnish, R. (1995), *Linguistics: An Introduction to Language and Communication*, 4th edn, Cambridge, MA: MIT Press.

Anderson, J. (1983), *The Architecture of Cognition*, Cambridge, MA: Harvard University Press.

Anderson, J. (1993), *Rules of the Mind*, Hillsdale, NJ: Lawrence Erlbaum.

Anderson, J. (1995), *Introduction to Neural Networks*, Cambridge, MA: MIT Press.

Anderson, J., and Bower, G. (1974), *Human Associative Memory*, New York: Hemisphere.

Anderson, J., and Rosenfeld, E. (eds) (1988), *Neurocomputing*, Cambridge, MA: MIT Press.

Anderson, J., and Rosenfeld, E. (eds) (1998), *Talking Nets: An Oral History of Neural Networks*, Cambridge, MA: MIT Press.

Angell, J. R. (1911), Usages of the terms mind, consciousness and soul, *Psychological Bulletin*, 8: 46-7.

Arbib, M. (ed.) (1995), *Handbook of Brain Theory*, Cambridge, MA: Bradford/MIT Press.

Aristotle, On memory and reminiscence. In R. McKeon (ed.) (1941), *The Collected Works of Aristotle*, New York: Random House.

Armstrong, D. M. (1999), *The Mind-Body Problem: An Opinionated Introduction*, Boulder, CO: Westview Press.

Aspry, W. (1990), *John von Neumann and the Origins of Modern Computing*, Cambridge, MA: MIT Press.

Aspry, W., and Burks, A. (eds) (1987), *Papers of John von Neumann on Computers and Computer Theory*, Cambridge, MA: MIT Press.

Atlas, J. (1997), On the modularity of sentence processing: semantical generality and the

language of thought. In J. Nuyts and E. Pederson (eds) (1997), *Language and Conceptualization*, Cambridge: Cambridge University Press.

Augarten, S. (1984), *Bit By Bit: An Illustrated History of Computers*, New York: Ticknor and Fields.

Baars, B. (1986), *The Cognitive Revolution in Psychology*, New York: Guilford Press.

Bain, A. (1855), *The Senses and the Intellect*, 4th edn, New York: D. Appleton (1902).

Bain, A. (1859), *The Emotions and the Will*, 4th edn, New York: D. Appleton.

Baker, L. Rudder (1987), *Saving Belief*, Princeton, NJ: Princeton University Press.

Ballard, D. (1997), *An Introduction to Neural Computation*, Cambridge, MA: Bradford/MIT Press.

Bara, B. (1995), *Cognitive Science: A Developmental Approach to the Simulation of the Mind*, Hillsdale, NJ: Lawrence Erlbaum.

Barr, A., Cohen, P., and Feigenbaum, E. (eds) (1981), *The Handbook of Artificial Intelligence*, vols 1-3, Los Altos, CA: William Kaufmann.

Barsalou, L. (1992), *Cognitive Psychology: An Overview for Cognitive Scientists*, Hillsdale, NJ: Lawrence Erlbaum.

Barwise, J., and Etchemendy, J. (1999), *Turing's World* 3.0: *An Introduction to Computability Theory*, Cambridge: Cambridge University Press.

Baumgartner, P., and Payr, S. (eds) (1995), *Speaking Minds: Interviews with Twenty Eminent Cognitive Scientists*, Princeton, NJ: Princeton University Press.

Beakley, B., and Ludlow, P. (eds) (1992), *The Philosophy of Mind: Classical Problems, Contemporary Issues*, Cambridge, MA: Bradford/MIT Press.

Beatty, J. (1995), *Principles of Behavioral Neuroscience*, Dubuque: Brown and Benchmark.

Bechtel, W. (1985), Contemporary connectionism: are the new parallel distributed processing models cognitive or associationist? *Behaviorism*, 13(1): 53-61.

Bechtel, W. (1994), Connectionism. In S. Guttenplan (ed.), *A Companion to the Philosophy of Mind*, Cambridge, MA: Blackwell.

Bechtel, W., and Abrahamsen, A. (1991), *Connectionism and the Mind*, Cambridge, MA: Blackwell.

Bechtel, W., and Graham, G. (eds) (1998), *A Companion to Cognitive Science*, Cambridge, MA: Blackwell.

Beckermann, A., Flohr, H., and Kim, J. (eds) (1992), *Emergence or Reduction? Essays on the Prospects of Nonreductive Physicalism*, New York: Walter de Gruyter.

Bever, T. (1992), The logical and extrinsic sources of modularity. In M. Gunnar and M. Maratsos (eds) (1992), *Modularity and Constraints in Language and Cognition*, Hillsdale, NJ: Lawrence Erlbaum.

Block, H. (1962), The Perceptron: a model for brain functioning. I, *Review of Modern Physics* 34: 123-35. Reprinted in Anderson and Rosenfeld (1988).

Block, N. (1978), Troubles with functionalism. In C. W. Savage (ed.) (1978), *Perception and Cognition: Issues in the Foundations of Psychology*, Minneapolis: University of Minnesota Press.

Block, N. (ed.) (1981), *Readings in the Philosophy of Psychology*, vol. 1, Cambridge, MA: Harvard University Press.

Block, N. (1983), Mental pictures and cognitive science, *Philosophical Review*, 90: 499-541.

Block, N. (1986), Advertisement for a semantics for psychology. In French et al. (1986).

Block, N. (1990), The computer model of the mind. In D. Osherson and E. Smith (eds) (1990), *An Invitation to Cognitive Science*, vol. 3: *Thinking*, Cambridge, MA: Bradford/MIT Press.

Block, N. (1994), Consciousness, Qualia. In S. Guttenplan (ed.) (1994), *A Companion to the Philosophy of Mind*, Cambridge, MA: Blackwell.

Block, N. (1995), On a confusion about a function of consciousness, *The Behavioral and Brain Sciences*, 18: 227-47. Reprinted in Block et al. (1997).

Block, N., Flanagan, O., and Guzeldere, G. (eds) (1997), *The Nature of Consciousness*, Cambridge, MA: Bradford/MIT Press.

Block, N., and Fodor, J. (1972), What psychological states are not, *Philosophical Review*, April: 159-81.

Block, N., and Segal, G. (1998), The philosophy of psychology. In A. C. Grayling (ed.) (1998), *Philosophy* 2, Oxford: Oxford University Press.

Bobrow, D., and Collins, A. (eds) (1975), *Representation and Understanding*, New York: Academic Press.

Bobrow, D., and Winograd, T. (1979), An overview of KRL, a Knowledge Representation Language, *Cognitive Science*, 1: 3-46.

Boden, M. (1977), Artificial Intelligence and Natural Man, New York: Basic Books.

Boden, M. (ed.) (1990), *The Philosophy of Artificial Intelligence*, Oxford: Oxford University Press.

Boolos G., and Jeffrey, R. (1989), *Computability and Logic*, 3rd edn, Cambridge: Cambridge University Press.

Boring, E. et al. (1939), *Introduction to Psychology*, New York: John Wiley.

Boring, E. (1951), *A History of Experimental Psychology*, 2nd edition, New York: Appleton, Century, Crofts.

Bower, G. (1975), Cognitive psychology: an introduction. In W. Estes (ed.), *Handbook of Learning and Cognitive Processes*, New York: Wiley.

Bower, G., and Clapper, J. (1989), Experimental Methods in cognitive science. In

Posner (1989).

Brachman, R. (1979), On the epistemological status of semantic networks. In Findler (1979).

Brachman, R., and Levesque, H. (eds) (1985), *Readings in Knowledge Representation*, San Mateo, CA: Morgan Kaufman.

Braddon-Mitchell, D., and Jackson, F. (1996), *Philosophy of Mind and Cognition*, Oxford: Blackwell.

Brand, M., and Harnish, R. (eds) (1986), *The Representation of Knowledge and Belief*, Tucson: University of Arizona Press.

Brink, F. (1951), Excitation and conduction in the neuron (ch. 2), Synaptic mechanisms (ch. 3). In S. S. Stevens, *Handbook of Experimental Psychology*, New York: Wiley.

Broadbent, D. (1957), A mechanical model for human attention and immediate memory, *Psychological Review*, 64: 205-15.

Broadbent, D. (1958), *Perception and Communication*, New York: Pergamon.

Bruner, J., Goodnow, J., and Austin, G. (1956), *A Study of Thinking*, New York: Wiley.

Burge, T. (1979), Individualism and the mental. In P. French et al. (eds) (1979), *Contemporary Perspectives in the Philosophy of Language*, Minneapolis: University of Minnesota Press.

Burks, A., Goldstine, H., and von Neumann, J. (1946), Preliminary discussion of the logical design of an electronic computing instrument. In Aspry and Burks (1987).

Cajal, R. y (1909), *New Ideas on the Structure of the Nervous System in Man and Vertebrates*, Cambridge, MA: Bradford/MIT Press.

Carnap, R. (1937), *The Logical Syntax of Language*, London: Routledge and Kegan Paul.

Carpenter, B., and Doran, R. (1977), The other Turing machine, *Computer Journal*, 20 (3): 269-79.

Carpenter, B., and Doran, R. (eds) (1986), *A. M. Turing's ACE Report of* 1946 *and Other Papers*, Cambridge, MA: MIT Press.

Carruthers, P., and Smith, P. (eds) (1996), *Theories of Theories of Mind*, Cambridge: Cambridge University Press.

Caudill, M., and Butler, C. (1993), *Understanding Neural Networks*, vols 1 and 2, Cambridge, MA: Bradford/MIT Press.

Chalmers, D. (1992), Subsymbolic computation and the Chinese room. In Dinsmore (1992).

Chalmers, D. (1995a), On implementing a computation, *Minds and Machines*, 4: 391-402.

Chalmers, D. (1995b), The puzzle of conscious experience, *Scientific American*, 273: 80-6.

Chalmers, D. (1995c), Facing up to the problem of consciousness, *Journal of Consciousness Studies*, 2(3): 200-19. Reprinted in Cooney (2000).

Chalmers, D. (1996a), Does a rock implement every finite-state automaton? *Synthese*, 108: 309-33.

Chalmers, D. (1996b), *The Conscious Mind*, Oxford: Oxford University Press.

Chater, N., and Oaksford, M. (1990), Autonomy, implementation and cognitive architecture: a reply to Fodor and Pylyshyn, *Cognition*, 34: 93-107.

Cherniak, E., and McDermott, D. (1985), *Introduction to Artificial Intelligence*, Reading, MA: Addison Wesley.

Chisley, R. (1995), Why everything doesn't realize every computation, *Minds and Machines*, 4: 403-20.

Chomsky, N. (1959), Review of Skinner, *Verbal Behavior*, *Language*, 35: 26-58.

Christensen, S., and Turner, D. (eds) (1993), *Folk Psychology and the Philosophy of Mind*, Hillsdale, NJ: Lawrence Erlbaum.

Church, A. (1936), An unsolvable problem of elementary number theory, *American Journal of Mathematics*, 58: 345-63. Reprinted in Davis (1965).

Churchland, P. M. (1981), Eliminative materialism and propositional attitudes, *Journal of Philosophy*, 78: 67-90.

Churchland, P. M. (1988), *Matter and Consciousness*, rev. edn, Cambridge, MA: Bradford/MIT Press.

Churchland, P. M. (1989), On the nature of explanation: a PDP approach. In P. M. Churchland (1989), *A Neurocomputational Perspective*, Cambridge, MA: Bradford/MIT Press. Reprinted with some changes in Haugeland (1997).

Churchland, P. M. (1990), Cognitive activity in artificial neural networks. In D. Osherson and E. Smith (eds) (1990), *An Invitation to Cognitive Science*, vol, 3, Cambridge, MA: Bradford/MIT Press.

Churchland, P. M. (1994), Folk psychology (2). In Guttenplan (1994).

Churchland, P. M., and Churchland, P. S. (1991), Could a machine think? *Scientific American*, 262(1): 32-7.

Churchland, P. S. (1986), *Neurophilosophy*, Cambridge, MA: Bradford/MIT Press.

Churchland, P. S., and Sejnowski, T. (1992), *The Computational Brain*, Cambridge, MA: Bradford/MIT Press.

Clark, A. (1989), *Microcognition*, Cambridge, MA: Bradford/MIT Press.

Clark, A. (1990), Connectionist minds, *Proceedings of the Aristotelian Society*. 83-102. Reprinted in MacDonald and Macdonald (1995).

Clark, A. (1991), In defense of explicit rules. In Ramsey, Stich, and Rumelhart (1991).

Clark, A. (1993), *Associative Engines*, Cambridge, MA: Bradford/MIT Press.

Clark, K., and Cowell, D. (1976), *Programs, Machines and Computability: An Introduction to the Theory of Computing*, New York: McGraw-Hill.

Clark, A., and Lutz, R. (eds) (1995), *Connectionism and Context*, Berlin: Springer Verlag.

Cognitive Science Society (eds), *Proceedings of the Annual Conference of the Cognitive Science Society*, Hillsdale, NJ: Lawrence Erlbaum (annual).

Cognitive Science 1978: Report of the State of the Art Committee to the Advisors of the Alfred P. Sloan Foundation, October 1, 1978. Reprinted in Machlup and Mansfield (1983).

Cohen, J., and Schooler, J. (eds) (1996), *Scientific Approaches to Consciousness*, Hillsdale, NJ: Lawrence Erlbaum.

Collins, A., and Smith, E. (eds) (1988), *Readings in Cognitive Science*, San Mateo, CA: Morgan Kaufman.

Cooney, B. (2000), *The Place of Mind*, Belmont, CA: Wadsworth.

Copeland, J. (1993a), The curious case of the Chinese gym, *Synthese*, 95: 173-86.

Copeland, J. (1993b), Artificial Intelligence: A Philosophical Introduction, Oxford: Biackwell.

Copeland, J. (1996a), What is computation? *Synthese*, 108: 335-59.

Copeland, J. (1996b), The Church-Turing thesis. In J. Perry and E. Zalta (eds), *The Stanford Encyclopedia of Philosophy*[http://plato. stanford. edu],

Copeland, J. (1997), The broad conception of computation, *American Behavioral Scientist*, 40(6): 690-716.

Copeland, J. (1998), Turing's O-machines, Searle, Penrose and the brain, *Analysis*, 58 (2): 128-38.

Copeland, J., and Proudfoot, D. (1996), On Alan Turing's anticipation of connectionism, *Synthese*, 108: 361-77.

Copeland, J., and Proudfoot, D. (1999), Alan Turing's forgotten ideas in computer science, *Scientific American*, April: 99-103.

Corsi, P. (1991), *The Enchanted Loom: Chapters in the History of Neuroscience*, Oxford: Oxford University Press.

Cowan, J., and Sharp, D. (1988), Neural nets and artificial intelligence. Reprinted in Graubard (1988).

Crane, T. (1995), *The Mechanical Mind*, New York: Penguin.

Crevier, D. (1993), *AI: The Tumultuous History of the Search for Artificial Intelligence*, New York: Basic Books.

Crick, F. (1994), *The Astonishing Hypothesis: The Scientific Search for the Soul*, New York: Touchstone/Simon and Schuster.

Crick, F., and Asanuma, C. (1986), Certain aspects of the anatomy and physiology of the cerebral cortex. In J. McClelland and D. Rumelhart (eds) (1986), *Parallel Distributed Processing*, vol. 2, Cambridge, MA: Bradford/MIT Press.

Cummins, R. (1986), Inexplicit information. In Brand and Harnish (1986).

Cummins, R. (1989), *Meaning and Mental Representation*, Cambridge, MA: Bradford/MIT Press.

Cummins, R. (1996), *Representations, Targets, and Attitudes*, Cambridge, MA: Bradford/MIT Press.

Cummins, R., and Cummins, D. (eds) (1999), *Minds, Brains and Computers: The Foundations of Cognitive Science*, Oxford: Blackwell.

Cussins, A. (1990), The connectionist construction of concepts. In Boden (1990).

D'Andrade, R. (1989), Cultural cognition. In Posner (1989).

Davies, M. (1991), Concepts, connectionism and the language of thought. In Ramsey, Stich, and Rumelhart (1991).

Davies, M. (1995), Reply: consciousness and the varieties of aboutness. In C. MacDonald and G. MacDonald (eds) (1995), *Philosophy of Psychology: Debates on Psychological Explanation*, Cambridge, MA: Blackwell.

Davis, M. (1958), *Computability and Unsolvability*, New York: McGraw-Hill.

Davis, M. (ed.) (1965), *The Undecidable*, Hewlett, NY: Raven Press.

Davis, S. (ed.) (1992), *Connectionism: Theory and Practice*, Oxford: Oxford University Press.

Dawson, M. (1998), *Understanding Cognitive Science*, Oxford: Blackwell.

Dennett, D. (1978a), Skinner skinned. In *Brainstorms*, Montgomery, VT: Bradford Books.

Dennett, D. (1978b), Towards a cognitive theory of consciousness. In *Brainstorms*, Montgomery, VT: Bradford Books.

Dennett, D. (1985), Can machines think? In M. Shafto (ed.) (1985), *How We Know*, New York: Harper and Row.

Dennett, D. (1987a), *The Intentional Stance*, Cambridge, MA: Bradford/MIT Press.

Dennett, D. (1987b), Fast thinking. In *The Intentional Stance*, Cambridge, MA: Bradford/MIT Press.

Dennett, D. (1991), *Consciousness Explained*, London: Penguin.

Dennett, D. (1995), *Darwin's Dangerous Idea*, London: Penguin.

Descartes, R. (1641), *Meditations.* In E. Haldane and G. Ross (1931), *Philosophical Works of Descartes*, vol. 1, Cambridge: Cambridge University Press. Reprinted by Dover Publications (1955).

Descartes, R. (1649), *Passions of the Soul.* In E. Haldane and G. Ross (1931), *Philosophical Works of Descartes*, vol. 1, Cambridge: Cambridge University Press,

Reprinted by Dover Publications (1955).

Devitt, M. (1989). A narrow representational theory of mind. In S. Silvers (ed.) (1989), *ReRepresentations*, Dordrecht: Kluwer.

Devitt, M. (1991), Why Fodor can't have it both ways. In Loewer and Rey (1991).

Devitt, M. (1996), *Coming to Our Senses*, Cambridge: Cambridge University Press.

Dinsmore, J. (ed.) (1992), *The Symbolic and Connectionist Paradigms*, Hillsdale, NJ: Lawrence Erlbaum.

Dretske, F. (1981), *Knowledge and the Flow of Information*, Cambridge, MA: Bradford/MIT Press.

Dretske, F. (1986a), Aspects of representation. In Brand and Harnish (1986).

Dretske, F. (1986b), Misrepresentation. In R. Bogdan (ed.) (1986), *Belief: Form. Content and Function*, Oxford: Oxford University Press.

Dretske, F. (1993), Conscious experience, *Mind*, 102: 263-83.

Dretske, F. (1995), *Naturalizing the Mind*, Cambridge, MA: Bradford/MIT Press.

Dreyfus, H. (1972/9), *What Computers Can't Do*, rev. edn, New York: Harper and Row.

Dreyfus, H., and Dreyfus, S. (1988), Making a mind vs. modeling the brain: artificial intelligence back at a branchpoint. Reprinted in Graubard (1988).

Dunlop, C., and Fetzer, J. (1993), *Glossary of Cognitive Science*, New York: Paragon House.

Ebbinghaus, H. (1885), *Memory: A Contribution to Experimental Psychology*, Columbia Teacher's College [1913].

Edelman, G. (1989), *The Remembered Present: A Biological Theory of Consciousness*, New York: Basic Books.

Elman, J. (1990a), Finding structure in time, *Cognitive Science*, 14: 213-52.

Elman, J. (1990b), Representation and structure in connectionist models. In G. Altman (ed.) (1990), *Cognitive Models of Speech Processing*, Cambridge, MA: Bradford/MIT Press.

Elman, J. (1992), Grammatical structure and distributed representations. In Davis (1992).

Elman, J. (1993), Learning and development in neural networks: the importance of starting small, *Cognition*, 48(1): 71-99.

Elman, J. et al. (1996), *Rethinking Innateness*, Cambridge, MA: Bradford/MIT Press.

Eysenck, M., and Keane, M. (1990), *Cognitive Psychology: A Student's Handbook*, Hillsdale, NJ: Lawrence Erlbaum.

Fancher, R. E. (1979), *Pioneers of Psychology*, New York: W. W. Norton.

Feigenbaum, E., and Feldman, E. (eds) (1963), *Computers and Thought*, New York: McGraw-Hill. Reissued (1995) Cambridge, MA: Bradford/MIT Press.

Feldman, J. (1989), Neural representation of conceptual knowledge. In Nadel et al. (1989).

Feldman, J., and Ballard, D. (1982), Connectionist models and their properties, *Cognitive Science*, 6: 205-54.

Field, H. (1977), Logic, meaning and conceptual role, *Journal of Philosophy*, July: 379-409.

Fikes, R., and Nilsson, N. (1971), STRIPS: a new approach to the application of theorem proving to problem solving, *Artificial Intelligence*, 2: 189-208.

Findler, N. (ed.) (1979), *Associative Networks*, New York: Academic Press.

Finger, S. (1994), *Origins of Neuroscience: A History of Explorations into Brain Function*, Oxford: Oxford University Press.

Flanagan, O. (1991), *The Science of the Mind*, 2nd edn, Cambridge, MA: Bradford/MIT Press.

Flanagan, O. (1992), *Consciousness Reconsidered*, Cambridge, MA: Bradford/MIT Press.

Fodor, J. (1975), *The Language of Thought*, Cambridge, MA: Harvard University Press.

Fodor, J. (1980a), Methodological solipsism considered as a research strategy in cognitive science, *The Behavioral and Brain Sciences*, 3(1): 63-73. Reprinted in Fodor (1981b).

Fodor, J. (1980b), Commentary on Searle, *The Behavioral and Brain Sciences*, 3: 431-2.

Fodor, J. (1981a), The mind-body problem, *Scientific American*, 244(1): 114-23.

Fodor, J. (1981b), *Representations*, Cambridge, MA: Bradford/MIT Press.

Fodor, J. (1983), *Modularity of Mind*, Cambridge, MA: Bradford/MIT Press.

Fodor, J. (1984), Semantics, Wisconsin style, *Synthese*, 59: 231-50. Reprinted in Fodor (1990).

Fodor, J. (1985), Precis of *The Modularity of Mind*, *The Behavioral and Brain Sciences*, 8: 1-6. Commentary on Fodor (1985), *The Behavioral and Brain Sciences*, 8: 7-42. Reprinted in Fodor (1990).

Fodor, J. (1986), The Modularity of Mind, and Modularity of Mind: Fodor's Response. In Z. Pylyshyn and W. Demopolous (eds) (1986), *Meaning and Cognitive Structure*, Norwood, NJ: Ablex.

Fodor, J. (1987), *Psychosemantics*, Cambridge, MA: Bradford/MIT Press.

Fodor, J. (1989), Why should the mind be modular? In G. Alexander (ed.), *Reflections on Chomsky*, Oxford: Blackwell. Reprinted in Fodor (1990).

Fodor, J. (1990), *A Theory of Content*, Cambridge, MA: Bradford/MIT Press.

Fodor, J. (1991), Replies. In Loewer and Rey (1991).

Fodor, J. (1994), *The Elm and the Expert*, Cambridge, MA: Bradford/MIT Press.

Fodor, J. (1998), *In Critical Condition*, Cambridge, MA: Bradford/MIT Press.

Fodor, J., and Lepore, E. (1991), Why meaning (probably) isn't conceptual role, *Mind and Language*, 4. Reprinted in S. Stich and T. Warfield (eds) (1994), *Mental Representation: A Reader*, Oxford: Blackwell.

Fodor, J., and Lepore, E. (1992), *Holism: A Shopper's Guide*, Cambridge, MA: Blackwell.

Fodor, J., and Pylyshyn, Z. (1988), Connectionism and cognitive architecture: a critical analysis. In S. Pinker and J. Mehler (eds), *Connections and Symbols*, Cambridge, MA: Bradford/MIT Press.

Forster, K. (1978), Accessing the mental lexicon. In E. Walker (ed.) (1978), *Explorations in the Biology of Language*, Cambridge, MA: Bradford/MIT Press.

Franklin, S., and Garzon, M. (1991), Neural computability. In O. Omidvar (ed) (1991) *Progress on Neural Networks*, vol. 1, Norwood, NJ: Ablex.

Frege, G. (1979), *Begriffsschrift*. Trans. T. Bynym (1979) as *Concept Script*, Oxford: Oxford University Press.

Frege, G. (1892), On sense and reference. Reprinted in Harnish (1994).

Frege, G. (1918), The thought: a logical inquiry, *Mind* 65: 289-311. Reprinted (with correction) in Harnish (1994).

French, P et al. (eds) (1986), *Midwest Studies in Philosophy*, X, Minneapolis: University of Minnesota Press.

French, R. (1990), Subcognition and the Turing test, *Mind*, 99: 53-65.

Gall, F (1835), *On the Functions of the Brain* and of Each of Its Parts, 6 volumes, Boston, MA: Marsh, Capen, and Lyon.

Gandy, R. (1988), The confluence of ideas in 1936. In R. Herken (ed.) (1988), *The Universal Turing Machine: A Half Century Survey*, Oxford: Oxford University Press.

Gardner, H. (1985), *The Mind's New Science*, New York: Basic Books.

Garfield, J. (ed.) (1987), Modularity in Knowledge Representation and Natural Language Understanding, Cambridge, MA: Bradford/MIT Press.

Garfield, J. (ed.) (1990), *Foundations of Cognitive Science*, New York: Paragon House.

Gazzaniga, M. (ed.) (1995), *The Cognitive Neurosciences*, Cambridge, MA: Bradford/MIT Press.

Gazzaniga, M., and LeDoux, J. (1978), *The Integrated Mind*, New York: Plenum Press.

Geschwind, N. (1979), Specialization of the human brain. In *The Brain*, San Francisco: Freeman.

Glymour, C. (1992), *Thinking Things Through*, Cambridge, MA: Bradford/MIT Press.

Gödel, K. (1931), [trans. from the original German as] "On formally undecidable propositions of Principia Mathematica and related systems". In Davis (1965).

Goel, V. (1992), Are computational explanations vacuous? *Proceedings of the 14th Annual Conference of the Cognitive Science Society*, Hillsdale, NJ: Lawrence Erlbaum.

Goldberg, S., and Pessin, A. (1997), *Gray Matters: An Introduction to the Philosophy of Mind*, Armonk, NY: M. E. Sharpe.

Goldman, A. I. (1993a), The psychology of folk psychology, *The Behavioral and Brain Sciences*, 16(1): 15-28.

Goldman, A. I. (1993b), Consciousness, folk psychology, and cognitive science, *Consciousness and Cognition*, 2: 364-82.

Goldman, A. I. (ed.) (1993c), *Readings in Philosophy and Cognitive Science*, Cambridge, MA: Bradford/MIT Press.

Goldman, A. I, (1993d), *Philosophical Applications of Cognitive Science*, Boulder, CO: Westview Press.

Goschke, T., and Koppelberg, S. (1991), The concept of representation and the representation of concepts in connectionist models. In Ramsey, Stich, and Rumelhart (1991).

Graubard, S. (ed.) (1988), The Artificial Intelligence Debate: False Starts, Real Foundations, Cambridge, MA: MIT Press.

Greenwood, J. (ed.) (1991), *The Future of Folk Psychology*, Cambridge: Cambridge University Press.

Gross, C., Rocha-Miranda, C., and Bender, D. (1972), Visual properties of neurons in the inferotemporal cortex of the macaque, *Journal of Neurophysiology*, 35: 96-111.

Gross, M., and Lentin, A. (1970), *Introduction to Formal Grammars*, New York: Springer-Verlag.

Guttenplan, S. (ed.) (1994), *A Companion to the Philosophy of Mind*, Oxford: Blackwetl.

Güzeldere, G. (1997), The many faces of consciousness: a field guide. In Block et al. (1997).

Haberland, K. (1994), *Cognitive Psychology*, Boston: Allyn and Bacon.

Halliday, M. A. K. (1970), Functional diversity in language as seen from a consideration of modality and mood in english, *Foundations of Language*, 6: 322-61.

Hameroff, S., et al. (eds) (1996), *Toward a Science of Consciousness: The First Tucson Discussions and Debates*, Cambridge, MA: Bradford/MIT Press.

Hanson, S. (1999), Connectionist neuroscience: representational and learning issues of

neuroscience. In Lepore and Pylyshyn (1999).

Harlow, H. (1953), Mice, monkeys, men and motives, *Psychological Review*, 60: 23-32.

Harman, G. (1982), Conceptual role semantics, *Notre Dame Journal of Formal Logic*, 23(2): 242-56.

Harman, G. (1987), (Non-solipsistic) conceptual role semantics. In Lepore (1987).

Harnish, R. (ed.) (1994), *Basic Topics in the Philosophy of Language*, Englewood Cliffs NJ: Prentice-Hall.

Harnish, R. (1995), Modularity and speech acts, *Pragmatics and Cognition*, 3(1): 2-29.

Harnish, R. (1996), Consciousness, cognitivism and computation: a reply to Searle, *Conceptus*, 29 Nr. 75: 229-49.

Harnish, R., and Farmer, A. (1984), Pragmatics and the modularity of the language system, *Lingua*, 63: 255-77.

Hartley, D. (1749), *Observations on Man*. Reprinted New York: Garland (1971).

Hatfield, G. (1992), Descartes' physiology and psychology. In J. Cottingham (ed.) (1992), *The Cambridge Companion to Descartes*, Cambridge: Cambridge University Press.

Haugeland, J. (ed.) (1981), *Mind Design*, Cambridge, MA: Bradford/MIT Press.

Haugeland, J. (1985), *Artificial Intelligence: The Very Idea*, Cambridge, MA: Bradford/MIT Press.

Haugeland, J. (ed.) (1997), *Mind-Design II*, Cambridge, MA: Bradford/MIT Press.

Hawking, S. (1988), *A Brief History of Time*, New York: Bantam Books.

Hayes, P. (1979), The naive physics manifesto. In D. Mitchie (ed.) (1979), *Expert Systems in the Electronic Age*, Edinburgh: Edinburgh University Press.

Hayes, P. (1980), The logic of frames. In D. Meting (ed.), *Frame Conceptions and Understanding*, Berlin: de Gruyter. Reprinted in Webber and Nilsson (1981).

Hebb, D. O. (1949), *The Organization of Behavior*, New York: Wiley.

Hebb, D. O. (1972), *A Textbook of Psychology*, 3rd edn, Toronto: W. B. Saunders.

Heims S. (1980), *John Neumann and Norbert Wiener*, Cambridge, MA: MIT Press.

Hernstein, R., and Boring, E. (eds.) (1966), *A Source Book in the History of Psychology*, Cambridge, MA: Harvard University Press.

Hewitt, C. (1971), Procedural embedding of knowledge in PLANNER. In *Proceedings of the Second Joint Conference on Artificial Intelligence*, pp. 167-82, London: British Computer Society.

Hewitt, C. (1990), The challenge of open systems. In D. Partridge and Y. Wilks (eds) (1990), *The Foundations of AI: A Sourcebook*, Cambridge: Cambridge University Press.

Hilgard, E. (1987), *Psychology in America*, New York: Harcourt Brace Jovanovich.

Hillis, D. (1985), *The Connection Machine*, Cambridge, MA: Bradford/MIT Press.

Hinton, G. (1992), How neural networks learn from experience, *Scientific American*, September.

Hirschfeld, L., and Gelman, S. (eds) (1994), *Mapping the Mind: Domain Specificity in Cognition and Culture*, Cambridge: Cambridge University Press.

Hirst, W. (ed.) (1988), *The Making of Cognitive Science: Essays in Honor of George A. Miller*, Cambridge: Cambridge University Press.

Hobbes, T. (1651), *Leviathan.* Reprinted Indianapolis: Bobbs-Merrill (1958).

Hodges, A. (1983), *Alan Turing: The Enigma*, New York: Simon and Schuster.

Hofstadter, D. (1979), *Gödel, Escher, Bach*, New York: Basic Books.

Hofstadter, D. (1981), The Turing test: a coffee house conversation. In D. Hofstadter and D. Dennett (eds) (1981), *The Mind's I*, New York: Basic Books.

Hopcroft, J., and Ullman, J. (1969), *Formal Languages and Their Relation to Automata*, New York: Addison Wesley.

Horgan, T. (1994), Computation and mental representation. In S. Stich and T. Warfield (eds) (1994), *Mental Representation*, Cambridge, MA: Blackwell.

Horgan, T., and Tienson, J. (eds) (1991), *Connectionism and the Philosophy of Mind*, Dordrecht: Kluwer Academic.

Hornik, K. et al. (1989), Multilayer feedforward networks are universal approximators, *Neural Networks*, 2: 359-66.

Horst, S. (1996), *Symbols, Computation and Intentionality; A Critique of the Computational Theory of Mind*, Berkeley: University of California Press.

Hubel, D., and Wiesel, T. (1979), Brain mechanisms of vision. In *The Brain*, San Francisco: Freeman.

Hume, D. (1739), *A Treatise of Human Nature*, Oxford: Oxford University Press (1880).

Hunt, M. (1993), *The Story of Psychology*, New York: Doubleday.

Ince, D. (ed.) (1992), *Mechanical Intelligence: Collected Works of A. M. Turing*, Amsterdam: North-Holland.

Jackson, F. (1986), What Mary didn't know, *Journal of Philosophy*, 83: 291-5.

Jacob, P. (1997), *What Minds Can Do*, Cambridge: Cambridge University Press.

James, W. (1890), *The Principles of Psychology*, New York: Dover.

James, W. (1892), *Psychology (Briefer Course)*, New York: Holt.

Jeffrey, R. (1991), *Formal Logic: Its Scope and Limits*, 3rd edn, New York: McGraw-Hill.

Johnson-Laird, P. (1988), *The Computer and the Mind: An Introduction to Cognitive Science*, Cambridge, MA: Harvard University Press.

Kandel, E., Schwartz, J., and Jessell, T. (1995), *Essentials of Neural Science and Behavior*, Norwalk: Appleton and Lange.

Karmiloff-Smith, A. (1992), *Beyond Modularity: A Developmental Perspective on Cognitive Science*, Cambridge, MA: Bradford/MIT Press.

Kent, E. (1981), *The Brains of Men and Machines*, Peterborough, NH: BYTE/McGraw-Hill.

Kilian, J., and Siegelmann, H. (1993), On the power of sigmoid neural networks, *Proceedings of the Sixth ACM Workshop on Computational Learning Theory*, 137-43. New York: Association for Computing Machinery.

Kim, J. (1993), *Supervenience and the Mind*, Cambridge: Cambridge University Press.

Kim, J. (1994), Supervenience. In Guttenplan (1994).

Kim, J. (1996), *Philosophy of Mind*, Boulder, CO: Westview.

Klahr, D., Langley, P., and Neches, R. (eds) (1987), *Production System Models of Learning and Development*, Cambridge, MA: Bradford/MIT Press.

Kobes, B. (1990), Individualism and artificial intelligence. In J. Tomberlin (ed.) (1990), *Philosophical Perspectives*, 4, Atascadero, CA: Ridgeview.

Kosslyn, S. (1980), *Image and Mind*, Cambridge, MA: Harvard University Press.

Kosslyn, S., and Koenig, O. (1992), *Wet Mind*, New York: Free Press.

Kurzweil, R. (1990), *The Age of Intelligent Machines*, Cambridge, MA: Bradford/MIT Press.

Kurzweil, R. (1999), *The Age of Spiritual Machines*, New York: Penguin Books.

Lackner, J., and Garrett, M. (1972), Resolving ambiguity: effects of biosing context in the unattended ear, *Cognition*, 1: 359-72.

Laird, J., Newell, A., and Rosenbloom, P. (1987), Soar: an architecture for general intelligence, *Artificial Intelligence*, 33(1): 1-64.

Lashley, K. (1923), The behavioristic interpretation of consciousness, *Psychological Review*, 30: 329-53.

Lashley, K. (1929), *Brain Mechanism and Intelligence*, Chicago: Chicago University Press.

Lashley, K. (1951), The problem of serial order in behavior. In L. Jeffress (ed.), *Cerebral Mechanisms in Behavior*, New York: Wiley.

Leahey, T. (1992), *A History of Psychology*, 3rd edn, Englewood Cliffs, NJ: Prentice-Hall.

Lehman, J., Laird, J., and Rosenbloom, P. (eds) (1998), A gentle introduction to Soar: an architecture for human cognition. In Scarborough and Sternberg (1998).

Lepore, E. (ed.) (1987), *New Directions in Semantics*, New York: Academic Press.

Lepore, E. (1994), Conceptual role semantics. In Guttenplan (1994).

Lepore, E., and Loewer, B. (1986), Solipsistic semantics. In French et al. (1986).

Lepore, E., and Pvlyshyn, Z. (eds) (1999), *What is Cognitive Science?*, Oxford: Blackwell.

Lettvin, J. et al. (1959), What the frog's eye tells the frog's brain. Reprinted in W. McCulloch (1965).

Levine, D. (1991), *Introduction to Neural Cognitive Modeling*, Hillsdale, NJ: Lawrence Erlbaum.

Levine, J. (1983), Materialism and qualia: the explanatory gap, *Pacific Philosophical Quarterly*, 64: 354-61.

Lindsay, P., and Norman, D. (1972), *Human Information Processing: An Introduction to Psychology*, New York: Academic Press.

Lisker, L., and Abramson, A. (1964), A cross-language study of voicing of initial stops: acoustical measurements, *Word*, 20: 384-422.

Loewer, B., and Rey, G. (eds) (1991), *Meaning in Mind*, Oxford: Blackwell.

Lormand, E. (1994), Qualia! (now playing at a theater near you), *Philosophical Topics*, 22(1 and 2): 127-56.

Luger, G. (ed.) (1995), *Computation and Intelligence*, Cambridge, MA: Bradford/MIT Press.

Lycan, W. (ed.) (1990), *Mind and Cognition*, Oxford: Blackwell.

Lycan, W. (1997), Consciousness as internal monitoring. In Block et al. (1997).

McClamrock, R. (1995), *Existential Cognition: Computational Minds in the World*, Chicago: University of Chicago Press.

McClelland, J. (1981), Retrieving general and specific information from stored knowledge of specifics. *Proceedings of the Third International Conference of the Cognitive Science Society*, Berkeley, 1981.

McClelland, J., and Rumelhart, D. (1981), An interactive activation model of context effects in letter perception: Part I. An account of basic findings, *Psychological Review*, 88(5): 375-407.

McClelland, J., and Rumelhart, D. (1988), *Explorations in Parallel Distributed Processing*, Cambridge, MA: Bradford/MIT Press.

McCorduck, P. (1979), *Machines Who Think*, San Francisco: W. H. Freeman.

McCulloch, G. (1995), *The Mind and Its World*, London: Routledge.

McCulloch, W. (1965), *Embodiments of Mind*, Cambridge, MA: MIT Press.

McCulloch, W., and Pitts, W. (1943), A logical calculus of the ideas immanent in nervous activity. Reprinted in W. McCulloch (1965), and in Anderson and Rosenfeld (1998).

McDermott, D. (1976), Artificial intelligence meets natural stupidity, SIGART Newsletter, 57. Reprinted in Haugeland (1981).

McDermott, D. (1986), A critique of pure reason. Research Report YALEU/CSD/RR

no. 480.

MacDonald, C., and MacDonald, G. (eds) (1995), *Connectionism: Debates on Psychological Explanation*, vol. 2, Oxford: Blackwell.

McGinn, C. (1982), The structure of content. In A. Woodfield (ed.) (1982), *Thought and Object*, Oxford: Oxford University Press.

McGinn, C. (1991), *The Problem of Consciousness*, Cambridge, MA: Blackwell.

Machlup, F., and Mansfield, U. (eds) (1983), *The Study of Information: Interdisciplinary Messages*, New York: Wiley.

McLaughlin, B. (1993), The connectionism/classicism battle to win souls, *Philosophical Studies*, 71:163-90.

McLaughlin, B., and Warfield, T. (1994), The allure of connectionism reexamined, *Synthese*, 101: 365-400.

McLeod, P., Plunkett, K., and Rolls, E. (1998), *Introduction to Connectionist Modelling of Cognitive Processes*, Oxford: Oxford University Press.

McTeal, M. (1987), *The Articulate Computer*, Oxford: Blackwell.

Maloney, J. C. (1989), *The Mundane Matter of the Mental Language*, Cambridge: Cambridge University Press.

Marcel, A., and Bisiach, E. (eds) (1988), *Consciousness in Contemporary Science*, Oxford: Oxford University Press.

Marr, D. (1977), Artificial intelligence — a personal view. Reprinted in Haugeland (1981).

Marr, D. (1982), *Vision*, San Francisco: Freeman.

Marx, M., and Hillix, W. (1963), *Systems and Theories in Psychology*, New York: McGraw-Hill.

Massaro, D. (1988), Some criticisms of connectionist models of human performance, *Journal of Memory and Language*, 27: 213-34.

Metropolis, N. et al. (eds) (1980), *A History of Computing in the Twentieth Century*, New York: Academic Press.

Mill, J. S. (1829), *The Analysis of the Phenomena of the Human Mind*, London: Baldwin and Cradock.

Mill, J. S. (1843), *A System of Logic*. Reprinted London: Longmans (1967).

Miller, G. (1951), *Language and Communication*, New York: McGraw-Hill.

Miller, G. (1956), The magical number seven, plus or minus two: some limits on our capacity for processing information, *Psychological Review*, 63: 81-97. Reprinted in Miller (1967).

Miller, G. (1967), *The Psychology of Communication*, New York: Basic Books.

Miller, G., Galanter, E., and Pribram, K. (1960), *Plans and the Structure of Behavior*, New York: Holt, Rinehart, and Winston.

Mills, S. (1990), Connectionism and eliminative materialism, *Acta Analytica*, 6: 19-31.

Minsky, M. (1966), Artificial intelligence. In *Information*, San Francisco: W. H. Freeman (originally published in *Scientific American*).

Minsky, M. (1967), *Computation: Finite and Infinite*, Englewood Cliffs, NJ: Prentice-Hall.

Minsky, M. (1975), A framework for representing knowledge. In P. Winston (ed.) (1975), *The Psychology of Computer Vision*, New York: McGraw-Hill. Reprinted in Collins and Smith (1988). Selections reprinted in Haugeland (1997).

Minsky, M., and Papert, S. (1969), *Perceptrons*, Cambridge, MA: MIT Press. Expanded edn 1988.

Moody, T. (1993), *Philosophy and Artificial Intelligence*, Englewood Cliffs: Prentice-Hall.

Morris, R. (1989), *Parallel Distributed Processing: Implications for Psychology and Neurobiology*, Oxford: Oxford University Press.

Mylopoulos, J., and Levesque, H. (1984), An overview of knowledge representation. In M. Brodie et al. (eds) (1984), *On Conceptual Modeling*, New York: Springer-Verlag.

Nadel, L., et al. (1986), The neurobiology of mental representation. In Brand and Harnish (1986).

Nadel, L., Cooper, L., Culicover, P., and Harnish, R. (eds) (1989), *Neural Connections, Mental Computation*, Cambridge, MA: Bradford/MIT Press.

Nagel, T. (1974), What is it like to be a bat?, *Philosophical Review*, 83: 435-50.

Nagel, T. (1993), What is the mind-body problem? In *Experimental and Theoretical Studies in Consciousness*, New York: Wiley.

Neisser, U. (1967), *Cognitive Psychology*, New York: Appleton, Century, Crofts.

von Neumann, J. (1945), First draft of a report on the EDVAC. In Aspry and Burks (1987).

Newell, A. (1973), Production systems: models of control structures. In W. Chase (ed.), *Visual Information Processing*, New York: Academic Press.

Newell, A. (1980), Physical symbol systems, *Cognitive Science*, 4(2): 135-83. Reprinted in D. Norman (ed.) (1981), *Perspectives in Cognitive Science*, Norwood, NJ: Ablex.

Newell, A. (1983), Reflections on the structure of an interdiscipline. In Machlup and Mansfield (1983).

Newell, A. (1990), *Unified Theories of Cognition*, Cambridge, MA: Harvard University Press.

Newell, A., Rosenbloom, P., and Laird, J. (1989), Symbolic structures for cognition. In Posner (1989).

Newell, A., and Simon, H. (1972), *Human Problem Solving*, Englewood Cliffs, NJ: Prentice-Hall.

Newell, A., and Simon, H. (1976), Computer science as an empirical inquiry, *Communications of the ACM*, 19(3): 113-26. Reprinted in Haugeland (1997).

Nietzsche, F. (1887), *On the Genealogy of Morals*, trans. (1967), New York: Vintage Books.

Nilsson, N. (1965), *Learning Machines*, New York: McGraw-Hill. Reissued with an introduction by T. Sejnowski as *The Mathematical Foundations of Learning Machines*, San Mateo, CA: Morgan Kaufman (1990).

Norman, D. (1981), What is cognitive science? In D. Norman (ed.) (1981), *Perspectives on Cognitive Science*, Norwood, NJ: Ablex.

Norman, D., and Rumelhart, D. (1975), *Explorations in Cognition*, San Francisco: Freeman.

Odifreddi, P. (1989), *Classical Recursion Theory*, Amsterdam: North Holland.

O'Keefe, J., and Nadel, L. (1978), *The Hippocampus as a Cognitive Map*, Oxford: Oxford University Press.

Osherson, D., et al. (eds) (1990, 1995), *An Invitation to Cognitive Science*, vols 1-3; (1998) vol. 4, Cambridge, MA: Bradford/MIT Press, 1st, 2nd edns.

Papert, S. (1988), One AI or many? Reprinted in Graubard (1988).

Partridge, D. (1996), Representation of knowledge. In M. Boden (ed.) (1996), *Artificial Intelligence*, New York: Academic Press.

Pavlov, I. (1927), *Conditioned Reflexes*, Oxford: Oxford University Press. Reprinted by Dover Books.

Pavlov, I. (1928), *Lectures on Conditioned Reflexes* (trans. W. Horsley Gant), New York: Liveright.

Penrose, R. (1989), *The Emperor's New Mind*, New York: Penguin Books.

Pessin, A, and Goldberg, S. (eds) (1996), *The Twin Earth Chronicles*, New York: M. E. Sharpe.

Plunkett, K., and Elman, J. (1997), *Exercises in Rethinking Innateness*, Cambridge, MA: Bradford/MIT Press.

Pohl, I., and Shaw, A. (1981), *The Nature of Computation*, Rockville, MD: Computer Science Press.

Posner, M. (ed.) (1989), *Foundations of Cognitive Science*, Cambridge,. MA: Bradford/MIT Press.

Post, E. (1943), Formal reductions of the general combinatorial problem, *American Journal of Mathematics*, 65: 197-268.

Putnam, H. (1960), Minds and machines. In S. Hook (ed.) (1960), *Dimensions of Mind*, New York: New York University Press. Reprinted in Putnam (1975a).

Putnam, H. (1967), The nature of mental states. Reprinted in Putnam (1975a). Originally titled "Psychological Predicates" and published in W. Capitain and D. Merrill (eds) (1967), *Art, Mind, and Religion*, Pittsburgh: University of Pittsburgh Press.

Putnam, H. (1975a), *Philosophical Papers*, vol. 2, Cambridge: Cambridge University Press.

Putnam, H. (1975b), The meaning of meaning. Reprinted in Putnam (1975a), and in Harnish (1994).

Putnam, H. (1981), Brains in a vat. In *Reason, Truth and History*, Cambridge: Cambridge University Press.

Putnam, H. (1988), *Representation and Reality*, Cambridge, MA: Bradford/MIT Press.

Pylyshyn, Z. (1979), Complexity and the study of artificial and human intelligence. Reprinted in Haugeland (1981).

Pylyshyn, Z. (1983), Information science: its roots and relations as viewed from the perspective of cognitive science. In Machlup and Mansfield (1983).

Pylyshyn, Z. (1984), *Computation and Cognition*, Cambridge, MA: Bradford/MIT Press.

Pylyshyn, Z. (1989), Computing in cognitive science. In Posner (1989).

Quinlan, P. (1966), *Semantic Memory*, Report AFCRL-66-189, Bolt Beranek and Newman, Cambridge, MA.

Quinlan, P. (1968), Semantic memory. In M. Minsky (ed.) (1986), *Semantic Information Processing*, Cambridge, MA: MIT Press. Reprinted in Collins and Smith (1988).

Quinlan, P. (1991), *Connectionism and Psychology*, New York: Harvester-Wheatsheaf.

Ramsey, W. (1992), Connectionism and the philosophy of mental representation. In S. Davis (ed.) (1992), *Connectionism: Theory and Practice*, Oxford: Oxford University Press.

Ramsey, W., Stich, S., and Garon, J. (1991), Connectionism, eliminativism, and the future of folk psychology. In Ramsey, Stich, and Rumelhart (1991). Also in Greenwood (1991). Reprinted in MacDonald and Macdonald (1995), and Stich (1996).

Ramsey, W., Stich, S., and Rumelhart, D. (eds) (1991), *Philosophy and Connectionist Theory*, Hillsdale, NJ: Lawrence Erlbaum.

Ratcliff, R. (1990), Connectionist models of recognition memory: constraints imposed by learning and forgetting functions, *Psychological Review*, 97(2): 285-308.

Rey, G. (1986), What's really going on in Searle's Chinese room?, *Philosophical Studies*, 50: 169-85.

Rey, G. (1997), *Contemporary Philosophy of Mind*, Oxford: Blackwell.

Rich, E. (1983), *Artificial Intelligence*, New York: McGraw-Hill.

Rolls, E., and Treves, A. (1998), *Neural Networks and Brain Function*, Oxford:

Oxford University Press.

Rosenberg, J. (1990a), Connectionism and cognition, *Acta Analytica*, 6: 33-46. Reprinted in Haugeland (1997).

Rosenberg, J. (1990b), Treating connectionism properly: reflections on Smolensky, *Psychological Research*, 52: 163-74.

Rosenblatt, E (1958), The perceptron: a probabilistic model for information storage and organization in the brain, *Psychological Review*, 63: 386-408. Reprinted in J. Anderson and E. Rosenfeld (eds) (1988).

Rosenblatt, F. (1962), *Principles of Neurodynamics*, Washington, DC: Spartan Books.

Rosenthal, D. (1986), Two concepts of consciousness, *Philosophical Studies*, 49: 329-59.

Rosenthal, D. (1997), A theory of consciousness. In Block et al. (1997).

Rumelhart, D. (1989), The architecture of mind: a connectionist approach. In Posner (1989). Reprinted in Haugeland (1997).

Rumelhart, D., and McClelland, J. (eds) (1986a), *Parallel Distributed Processing*, vols 1 and 2, Cambridge, MA: Bradford/MIT Press.

Rumelhart, D., and McClelland, J. (1986b), On learning the past tenses of English verbs. In Rumelhart and McClelland (1986a), vol. 2, ch. 17.

Rumelhart, D., and Zipser, D. (1986), Feature discovery by competitive learning. In Rumelhart and McClelland (1986a), vol. 1, ch. 5.

Russell, B. (1918), The philosophy of logical atomism, *The Monist*. Reprinted (1985) La Salle, IL: Open Court.

Scarborough, D., and Sternberg, S. (eds) (1998), *An Invitation to Cognitive Science*, vol. 4, Cambridge, MA: Bradford/MIT Press.

Schank, R., and Abelson, R. (1977), *Scripts, Plans, Goals and Understanding*, Hillsdale, NJ: Lawrence Erlbaum. Chs 1-3 reprinted in Collins and Smith (1988).

Schneider, W. (1987), Connectionism: is it a paradigm shift for psychology?, *Behavior Research Methods, Instruments, and Computers*, 19: 73-83.

Schwartz, J. (1988), The new connectionism: developing relations between neuroscience and artificial intelligence. In Graubard (1988).

Scott, D. (1967), Some definitional suggestions for automata theory, *Journal of Computer and System Sciences*, 1: 187-212.

Seager, W. (1999), *Theories of Consciousness*, New York: Routledge.

Searle, J. (1969), *Speech Acts*, Cambridge: Cambridge University Press.

Searle, J. (1979), What is an intentional state? *Mind*, January: 74-92.

Searle, J. (1980), Minds, brains and programs, *Behavioral and Brain Sciences*, 3: 417-24. Reprinted in Haugeland (1997).

Searle, J. (1983), *Intentionality*, Cambridge: Cambridge University Press.

Searle, J. (1990a), Consciousness, explanatory inversion, and cognitive science, *Behavioral*

and Brain Sciences, 13(4): 585-642.

Searle, J. (1990b), Is the brain a digital computer?, *Proceedings of the American Philosophical Association*, 64(3): 21-37. Incorporated into Searle (1992), ch. 9.

Searle, J. (1991), Is the brain's mind a computer program?, *Scientific American*, 262 (1): 26-31.

Searle, J. (1992), *The Rediscovery of the Mind*, Cambridge, MA: Bradford/MIT Press.

Searle, J. (1997), *The Mystery of Consciousness*, New York: New York Review.

Sechenov, L. (1863), *Reflexes of the Brain*. Excerpted in R. Hernstein and E. Boring (eds) (1965), *A Source Book in the History of Psychology*, Cambridge, MA: Harvard University Press.

Segal, G. (1996), The modularity of theory of mind. In P. Carruthers and P. Smith (eds) (1996), *Theories of Mind*, Cambridge: Cambridge University Press.

Seidenberg, M. , et al. (1982), Automatic access to the meanings of ambiguous words in context, *Cognitive Psychology*, 14: 489-537.

Sejnowski, T. , and Rosenberg, C. (1987), Parallel networks that learn to pronounce English text, *Complex Systems*, 1: 145-68.

Selfridge, O. (1959), Pandemonium: a paradigm for learning. In D. Blake et al. (eds), *Proceedings of the Symposium on the Mechanization of Thought Processes*, National Physical Laboratory, London: HMSO.

Selfridge, O. , and Neisser, U. (1960), Pattern recognition by machine, *Scientific American*, August: 60-8.

Shannon, C. , and Weaver, W. (1949), *The Mathematical Theory of Communication*, Urbana: University of Illinois Press.

Shear, J. (ed.) (1997), *Explaining Consciousness: The Hard Problem*, Cambridge, MA: Bradford/MIT Press.

Shepherd, G. (1991), *Foundations of the Neuron Doctrine*, Oxford: Oxford University Press.

Shepherd, G. (1994), *Neurobiology*, 3rd edn, Oxford: Oxford University Press.

Sherrington, C. (1906), *The Integrative Action of the Nervous System*, reprinted (1973), New York: Arno Press.

Shortliff, E. H. (1976), *Computer-based Medical Consultants: MYCIN*, New York: North-Holland.

Siegelmann, H. , and Sontag, E. (1992), On the computational power of neural nets, *Proceedings of the Fifth ACM Workshop on Computational Learning Theory*, 440-449.

Skinner, B. F. (1938), *The Behavior of Organisms*, Englewood Cliffs, NJ: Prentice-Hall.

Skinner, B. F. (1957), *Verbal Behavior*, Englewood Cliffs, NJ: Prentice-Hall.

Skinner, B. F. (1976), *About Behaviorism*, New York: Knopf.

Smith, L. (1986), *Behaviorism and Logical Positivism*, Stanford, CA: Stanford University Press.

Smolensky, P. (1988a), On the proper treatment of connectionism, *Behavioral and Brain Sciences*, 11: 1-23. *Peer commentary*. 23-58. *Author's replies*: 59-74.

Smolensky, P. (1988b), *Lectures on Connectionist Cognitive Modeling* (unpublished manuscript).

Smolensky, P. (1989), Connectionist modelling: neural computation, mental connections. In Nadel et al. (1989). Reprinted in Haugeland (1997).

Smolensky, P. (1990), Representation in connectionist networks. In D. Memmi and Y. Visetti (eds) (1990), *Intellectica*(Special issue 9-10) *Modeles Connexionists*.

Smolensky, P. (1991a), Connectionism, constituency, and the language of thought. In Loewer and Rey (1991).

Smolensky, P. (1991b), The constituent structure of connectionist mental states: a reply to Fodor and Pylyshyn. In Morgan and Tienson (1991).

Smolensky, P. (1995), On the projectable predicates of connectionist psychology: a case for belief. In MacDonald and MacDonald (1995).

Squire, L., and Kosslyn, M. (eds) (1998), *Findings and Current Opinion in Cognitive Neuroscience*, Cambridge, MA: Bradford/MIT Press.

Stabler, E. (1983), How are grammars represented?, *Behavioral and Brain Sciences*, 6: 391-421.

Staugaard, Jr., A. (1987), *Robotics and AI*, Englewood Cliffs, NJ: Prentice-Hall.

Sterelny, K. (1990), *The Representational Theory of Mind*, Oxford: Blackwell.

Sternberg, S. (1970), Memory-scanning: mental processes revealed by reaction-time experiments, *American Scientist*, 57: 421-57.

Stich, S. (1978), Autonomous psychology and the belief desire thesis, *The Monist*, 61: 573-90.

Stich, S. (1983), *From Folk Psychology to Cognitive Science*, Cambridge, MA: Bradford/MIT Press.

Stich, S. (1991), Narrow content meets fat syntax. In Loewer and Rey (1991).

Stich, S. (1996), *Deconstructing the Mind*, Oxford: Oxford University Press.

Stich, S., and Warfield, T. (1995), Reply to Clark and Smolensky: do connectionist minds have beliefs? In MacDonald and MacDonald (1995).

Stillings, N., et al. (1995), *Cognitive Science: An Introduction*, 2nd edn, Cambridge, MA: Bradford/MIT Press.

Stone, J. (1972), Morphology and physiology of the geniculo-cortical synapse in the cat: the question of parallel input to the striate cortex, *Investigative Ophthalmology*, 11, 338-46.

Sutherland, N. S. (ed.) (1989), *The International Dictionary of Psychology*, New

York: Continuum.

Tarski, A. (1969), Truth and proof, *Scientific American*, June: 63-77.

Tennant, H. (1981), *Natural Language Processing*, New York: Petrocelli.

Thagard, P. (1986), Parallel computation and the mind-body problem, *Cognitive Science*, 10: 301-18.

Thagard, P. (1996), *Mind: An Introduction to Cognitive Science*, Cambridge, MA: Bradford/MIT Press.

Thagard, P. (ed.) (1998), *Mind Readings*, Cambridge, MA: Bradford/MIT Press.

Thorndike, E. (1911), *Animal Intelligence*, New York: Hafner.

Thorndike, E. (1929), *Human Learning*, New York: Johnson Reprint Corporation.

Tienson, J. (1991), Introduction. In Horgan and Tienson (1991).

Tremblay, J.-P., and Sorensen, P. (1984), *An Introduction to Data Structures and Their Applications*, 2nd edn, New York: McGraw-Hill.

Turing, A. (1936/7), On computable numbers, with an application to the *entscheidungsproblem*, *Proceedings of the London Mathematical Society*, Series 2, 42: 230-65. Reprinted in Davis (1965).

Turing, A. (1946), Proposal for development in the mathematics division of an automatic computing engine (ACE). Reprinted in Carpenter and Doran (1986).

Turing, A. (1948), Intelligent machinery. In B. Meltzer and D. Michie (eds) (1969), *Machine Intelligence*, 5: 3-23, New York: Academic Press. Reprinted in Ince (1992).

Turing, A. (1950), Computing machinery and intelligence, *Mind*, 236. Reprinted in Haugeland (1997), and in Ince (1992).

Valentine, E. (1989), Neural nets: from Hartley and Hebb to Hinton, *Journal of Mathematical Psychology*, 33: 348-57.

van Gelder, T. (1991), What is the "D" in "PDP"? A survey of the concept of distribution. In Ramsey, Stich, and Rumelhart (1991).

van Gelder, T. (1992), The proper treatment of cognition. In *Proceedings of the 14th Annual Conference of the Cognitive Science Society*, Hillsdale, NJ: Erlbaum.

van Gelder, T. (1997), Dynamics in cognition. In Haugeland (1997).

Verschure, P. (1992), Taking connectionism seriously. In *Proceedings of the 14th Annual Conference of the Cognitive Science Society*, Hillsdale, NJ: Erlbaum.

von Eckhardt, B. (1993), *What is cognitive science*? Cambridge, MA: Bradford/MIT Press.

von Eckhardt, B. (1994), Folk psychology (1). In Guttenplan (1994).

Walker, S. F. (1990), A brief history of connectionism and its psychological implications, *AI and Society*, 4: 17-38.

Waltz, D. (1982), Artificial intelligence, *Scientific American*, October.

Wang, H. (1957), A variant of Turing's theory of calculating machines, *Journal of the*

Association of Computing Machinery, 4: 63-92.

Warner, H. (1921), *A History of Association Psychology*, New York: Charles Scribner's Sons.

Wasserman, P. (1989), *Neural Computing: Theory and Practice*, New York: Van Nostrand Reinhold.

Watson, J. B. (1913), Psychology as the behaviorist views it, *Psychological Review*, 20: 158-77.

Webber, B., and Nilsson, N. (eds) (1981), *Readings in Artificial Intelligence*, San Mateo, CA: Morgan Kaufman.

Weiskrantz, L. (1988), Some contributions of neuropsychology of vision and memory to the problem of consciousness. In A. Marcel and E. Bisiach (eds) (1988), *Consciousness in Contemporary Science*, Oxford: Oxford University Press.

White, R. (2000), Some basic concepts of computability theory. In B. Cooney (ed.), *The Place of Mind, Belmont, CA: Wadsworth.*

Whitehead, A. N., and Russell, B. (1910-13), *Principia Mathematica*, vols 1-3, Cambridge: Cambridge University Press. 2nd edn (1925).

Wilks, Y. (1977), Natural language understanding systems with the A. I. paradigm: a survey and some comparisons. In Zampoli (1977).

Wilson, F. (1992), Association, ideas, and images in Hume. In P. Cummins and G. Zoeller (eds) (1992), *Minds, Ideas and Objects*, Atascadero, CA: Ridgeview.

Wilson, R., and Keil, F. (eds) (1999), *The MIT Encyclopedia of the Cognitive Sciences*, Cambridge, MA: MIT Press.

Winograd, T. (1972), *Understanding Natural Language*, New York: Academic Press.

Winograd, T. (1973), A procedural model of language understanding. In R. Shank and K. Colby (eds) (1973), *Computer Models of Thought and Language*, San Francisco: Freeman.

Winograd, T. (1975), Frame representations and the declarative/procedural controversy. In Bobrow and Collins (1975).

Winograd, T. (1977), Five lectures on artificial intelligence. In Zampolli (1977).

Winograd, T. (1980), What does it mean to understand language?, *Cognitive Science*, 4: 209-41.

Winograd, T. (1983), *Language as a Cognitive Process*, Reading, MA: Addison-Wesley.

Winston, P. (1977), *Artificial Intelligence*, Reading, MA: Addison Wesley.

Woods, W. (1975), What's in a link? In Bobrow and Collins (1975). Reprinted in Collins and Smith (1988).

Young, R. M. (1970), *Mind, Brain and Adaption in the Nineteenth Century*, Oxford: Oxford University Press.

Zampoli, A. (ed.) (1977), *Linguistic Structures Processing*, New York: North-Holland.

索　引*

* 索引中所标注的页码为原著页码。——译者注

E

F

G

H

K

L

M

N

O

Q

T

U

译后记

本书是一本对认知科学基础作跨学科和历史性评述的著作。该书整合了认知科学的广义(研究认知现象的一门交叉学科)和狭义(研究心智表征和计算能力的学科)两种理解,将之概括为:“认知科学的研究对象是认知,认知科学的方法是组成认知科学各门学科自身的方法,该领域的核心假设为心理状态和过程可计算”。其内容也即围绕着“认知可计算”这一核心假设而展开,将“计算”作为连接“心智、大脑与计算机”三者的基础。书中第一部分,以认知科学重要的软件(心智)和硬件(脑)隐喻为线索,概述了认知科学发展的有关哲学、生理学、心理学和计算机科学方面的历史背景。作者认为,麦卡洛克(W. McCulloch)和皮茨(W. Pitts)《神经活动内在概念的逻辑演算》一文的出现标志着软件、硬件两个主题开始走向融合(计算机),促进了心智计算理论(CTM)的发展,并且也是 CTM 的两种具体理论——心智数字计算理论(DCTM)和心智联结计算理论(CCTM)的分歧点。第二、三部分,从历史的视角分析了若干具有代表性的认知模型,分别对 DCTM 和 CCTM 的发展及其在结构、操作及表征等基础概念上的特征进行了详细评述,并指出了它们各自所具备的优势和不足。

本书从历史的视角概述了认知科学的发展过程,其中涉及了多个学科有关认知的重要议题,如哲学、心理学、计算机科学、神经生理学、语言学等。全书结构合理,内容丰富,每一章都提供了相关思考题和推荐读物,对于希望较为深入和全面地了解和认识认知科学的基础性概念、理论及历史演变过程的读者来说,无疑是一本理想的导论性教材。正如马里兰大学哲学教授 Georges Rey 所言:“本书不仅对于研习认知科学的学生非常重要,而且对那些想要获取认知科学内自己从事领域之外其他领域知识的学者,以及想

对这门学科的发展历史有深入了解的人，都会从中获益。”

参与本书部分章节早期试译工作的研究生包括：严密、吴燕、董彦、朱健、刘美平、王琎、阮奔奔、秦艳燕、杜莹、吴依云。提供本书初译稿的研究生分别是：李鹏鑫（导论、第1—6、10、11章、结语）、韩玮（第7章）、邹慧民（第8章）、陈刚（第9章）、赵明（第12章）、张亚萍（第13章）。全书由王淼、李鹏鑫负责统译、修改和审校。在翻译和审校过程中，我们得到了唐孝威院士、黄华新教授、盛晓明教授、罗见今教授、李恒威副教授给予的热情支持、鼓励与帮助。原书作者哈尼什（R. M. Harnish）教授关心本书的翻译和出版工作，并应译者之邀专门撰写了中文版序言。浙江大学出版社对本书的出版给予了大力支持和帮助，在此一并表示感谢。

由于译者的水平和知识面有限，译文中难免存在疏漏与不妥之处，敬请专家和读者批评指正。

译　者

2010年6月

[General Information]
书名=心智、大脑与计算机　人是科学创立史导论
作者=（美）R.M.哈尼什著
页数=418
SS号=12714102
出版日期=2010.08